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 $\mathbb{R}^n$?

Solution:

1. VC dimension is 6. It is straightforward to see that the following set of six points can be shattered:

$$\{(1,0,0),(-1,0,0),(0,1,0),(0,-1,0),(0,0,1),(0,0,-1)\}.$$

For any seven points $x_1 = (x_1^1, x_2^1, x_3^1), \ldots, x_7 = (x_7^1, x_7^2, x_7^3)$, let $x_{\min}^j = \min_{i=1}^7 x_i^j$ and $x_{\max}^j = \max_{i=1}^7 x_i^j$ for any $j \in \{1,2,3\}$. We define hyper-rectangle $R = \left[x_{\min}^1, x_{\max}^1\right] \times \left[x_{\min}^2, x_{\max}^2\right] \times \left[x_{\min}^3, x_{\max}^3\right]$. Therefore, either a single point resides within the interior of R, or there exist two points sharing an identical boundary point. Thus, it is impossible to shatter seven points, thereby establishing the VC-dimension at 6.

2. VC dimension is 4. It is straightforward to see that any four points can be shattered. However, for a sequence of five points on a line, it is not possible to shatter them if successive points are labeled with alternating labels, starting with a positive label. Thus, the VC-dimension of subsets of the real line formed by the union of two intervals is 4

3. VC dimension is 1 or 2. If we only use spheres to include or exclude points normally, we can only fully separate one point. So, the VC-dimension is 1. If we add a way to flip the classification, we can fully separate two points in any way we want. This makes the VC-dimension 2.

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:

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

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:

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

(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 $i \in [k]$,

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

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

Solution:

(a) First split the term into two separate terms:

$$\begin{aligned}|L_S(h_1) - L_S(h_2)| &\leq |R(h_1) - R(h_2)| + |\widehat{R}_S(h_1) - \widehat{R}_S(h_2)| \\ &= \Big| \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}}[(h_1(x) - y)^2 - (h_2(x) - y)^2] \Big| + \Big| \frac{1}{m} \sum_{i=1}^{m} (h_1(x_i) - y_i)^2 - (h_2(x_i) - y_i)^2 \Big|.\end{aligned}$$

Then, expanding the following term,

$$\begin{aligned}(h_1(x) - y)^2 - (h_2(x) - y)^2 &= (h_1(x) - h_2(x))(h_1(x) + h_2(x) - 2y) \\ &= (h_1(x) - h_2(x))\big((h_1(x) - y) + (h_2(x) - y)\big) \leq \|h_1 - h_2\|_\infty 2M \,,\end{aligned}$$

allows us to bound both the empirical and true error, resulting in a total bound of $4M\|h_1 - h_2\|_\infty$.

(b) This follows by splitting the event into the union of several smaller events and then using the sum rule,

$$\mathop{\mathbb{P}}_{S}\Big[\sup_{h \in \mathcal{H}} |L_S(h)| \geq \epsilon\Big] = \mathop{\mathbb{P}}_{S}\Big[\bigvee_{i=1}^{k} \sup_{h \in B_i} |L_S(h)| \geq \epsilon\Big] \leq \sum_{i=1}^{k} \mathop{\mathbb{P}}_{S}\Big[\sup_{h \in B_i} |L_S(h)| \geq \epsilon\Big].$$

(c) For any $h \in B_i$, we have $|L_S(h) - L_S(h_i)| \leq 4M\|h - h_i\|_\infty \leq \epsilon/2$. Thus, if for any $h \in B_i$ we have $|L_S(h)| \geq \epsilon$ it must be the case that $|L_S(h_i)| \geq \epsilon/2$, which shows the inequality.
To complete the bound, we use Hoeffding's inequality applied to the random variables $(h(x_i) - y_i)^2/m \leq M^2/m$, which guarantees

$$\mathop{\mathbb{P}}_{S}\Big[|L_S(h_i)| \geq \frac{\epsilon}{2}\Big] \leq 2\exp\Big(\frac{-m\epsilon^2}{2M^4}\Big).$$

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.

Solution:

(a) Let $S = \{x_1, x_2, \ldots, x_m\}$ be a ϵ-packing set in F. By definition, for any $x_i, x_j \in S$ with $i \neq j$, $d(x_i, x_j) > \epsilon$. Consider ϵ-balls centered at each point $x_i \in S$. By the maximality of the ϵ-packing set S, adding any additional point to S would violate the condition $d(x_i, x_j) > \epsilon$. This implies that any point in F must be within ϵ distance of some point in S. Thus, F can be covered by ϵ-balls centered at points in S, meaning:

$$\mathcal{N}(F,\epsilon) \leq |S| = \mathcal{M}(F,\epsilon).$$

(b) Let V = volume(unit ball). Assume that $r > \epsilon$. Then you can observe that the $\frac{\epsilon}{2}$-balls centered at the members of any ϵ-packing of a ball are disjoint. So, the balls are included in balls of radius $r + \frac{\epsilon}{2}$. Thus, $\mathcal{M}(B,\epsilon)\frac{\epsilon}{2}^d V \leq (r + \frac{\epsilon}{2})^d V \leq \frac{3r}{2}^d V$. This implies that $\mathcal{M}(B,\epsilon) \leq (3r/\epsilon)^d$.

(c) This should be clear since an $\frac{\epsilon}{\mu}$ cover of the unit ball in $\mathbb{R}^d$ for norm $\|\cdot\|$ provides an ϵ cover of G: $\mathcal{N}_\infty(G,\epsilon) \leq \mathcal{N}_{\|\cdot\|}(B(0,1), \frac{\epsilon}{\mu})$, which leverages the definition of μ-Lipschitz continuity. The bound results from the previous result.

(d) This should be clear by directly applying the bounds established in Problem B.