ECE 434: Problem Set 2

Due on October 26, 2023

Professor Chi Jin

Evan Dogariu
Collaborators: Alex Zhang

Problem 1

Compute the exact VC dimension for the following boolean classes.

(a) Indicator functions on all half-spaces in $\mathbb{R}^d$:

$$\mathcal{F} := \left\{ f(x) = \mathbb{1}_{\{w^T x \leq t\}} : \; w \in r^d, \; t \in \mathbb{R} \right\}$$

(b) Let $\mathcal{F}$ be the collection of all polytopes in $\mathbb{R}^2$, where we define a polytope in the plane as a convex hull of a collection of finitely many points.

Solution

Proof. (a) We first show that $\mathcal{F}$ can shatter a collection of $d+1$ points. In particular, let $\{e_j\}_{j=1}^d$ denote the standard basis, and consider the collection $\{0, e_1, \ldots, e_d\}$ with 0 the origin. Let $(y_0, y_1, \ldots, y_d) \in \{0,1\}^{d+1}$ be any arbitrary labeling of these points. We construct the vector w by saying that for $j = 1, \ldots, d$, let

$$w_j := \begin{cases} 1 & y_j = 1 \\ -1 & y_j = 0 \end{cases}$$

Then, $w^T e_j = w_j$. Let $t := \begin{cases} 0 & y_0 = 1 \\ -0.1 & y_0 = 0 \end{cases}$. Then, we have $w^T e_j = w_j \leq t \iff y_j = 1$ and also $w^T 0 = 0 \leq t \iff y_0 = 1$, correctly classifying these points with these labels. Since this holds for all labelings, we have shattered these $d+1$ points. So, the VC dimension is $\geq d+1$.

Suppose now that we have an arbitrary collection of $d+2$ points, say $\{x_1, \ldots, x_{d+2}\}$. Note that we may lift to a higher dimension with homogenous coordinates, and so the function space

$$\tilde{\mathcal{F}} := \left\{ f(x) = \mathbb{1}_{\tilde{w}^T [x,1] \leq 0} : \; \tilde{w} \in \mathbb{R}^{d+1} \right\}$$

is equal to $\mathcal{F}$ (here, we use the notation $[x, 1]$) to describe concatenating 1 to the end of the vector x. $\tilde{\mathcal{F}} = \mathcal{F}$ since for any $\tilde{w} = [w, t] \in \mathbb{R}^{d+1}$ (here, $t \in \mathbb{R}$ and $w \in \mathbb{R}^d$) we know that $\tilde{w}^T [x, 1] \leq 0 \iff w^T x + t \leq 0 \iff w^T x \leq t$, and so for each function in one of the hypothesis classes we may find a corresponding function in the other. Define $\widetilde{x_j} := [x_j, 1]$. Now, we know that since there are $d+2$ points $\widetilde{x_j} \in \mathbb{R}^{d+1}$, one of the points must be linearly dependent on the others. As such, there exist α_j's in $\mathbb{R}$ (at least one of which is nonzero) for which

$$\widetilde{x_k} = \sum_{j \neq k} \alpha_j \widetilde{x_j}$$

We will construct a labeling of the points that cannot be classified as follows: set

$$y_j = \begin{cases} 0 & j = k \text{ or } \alpha_j \leq 0 \\ 1 & j \neq k \text{ and } \alpha_j > 0 \end{cases}$$

Now, suppose by way of contradiction that some $\tilde{w} \in \mathbb{R}^{d+1}$ correctly classifies these points with this labeling. Therefore, for all $j \neq k$ we have that $\widetilde{w}^T \widetilde{x_j} \leq 0 \iff \alpha_j > 0$. So, looking at the k^{th} point,

$$\widetilde{w}^T \widetilde{x_k} = \widetilde{w}^T \left(\sum_{j \neq k} \alpha_j \widetilde{x_j} \right) = \sum_{j \neq k} \alpha_j \widetilde{w}^T \widetilde{x_j}$$

We know that each $\alpha_j \widetilde{w}^T \widetilde{x_j} \leq 0$ since they have opposite signs. Therefore, $\widetilde{w}^T \widetilde{x_k} \leq 0$ as well. However, we set $y_k = 0$, which would require that $\widetilde{w}^T \widetilde{x_k} > 0$ for correct classification. This is a contradiction, and so $\widetilde{w}$

Problem 1 continued on next page...

cannot correctly classify these points with these labels. Since this holds for all labelings of all collections of $d+2$ distinct points, the VC dimension is $< d+2$.

(b) We claim that the VC dimension of $\mathcal{F}$ is infinite. To see this, let $n \in \mathbb{N}$ be arbitrary; we will show we can shatter n points. Let $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ be n distinct points distributed around the unit circle (say, the n^{th} roots of unity). Consider any arbitrary labeling $(y_1, \ldots, y_n) \in \{0,1\}^n$ of these points (if less than 3 vertices are labeled 1 we can classify it trivially by infinitesimally encircling the line connecting the two 1-labeled points, infinitesimally encircling the only 1-labeled point, or drawing a far away polytope if $y_j = 0 \, \forall j$; so suppose WOLOG that at least 3 points have label 1). Define E to be the convex polygon formed by the convex hull of the points $\{x_j : y_j = 1\}$ (such that E contains its vertices). Then, $\mathbb{1}_E \in \mathcal{F}$ by definition. Also, since E contains its vertices, we know $y_j = 1 \implies x_j \in E$. Now, suppose that j is such that $y_j = 0$. Then, x_j cannot lie in E; to see this, consider the two points x_ℓ and x_r closest to x_j clockwise and counterclockwise for which $y_\ell = y_r = 1$. Then, the line connecting x_ℓ and x_r forms an edge of E by construction. However, the lines $x_j - x_\ell$ and $x_j - x_r$ are supporting hyperplanes of E, and so they are tangent to the boundary of E exactly at x_ℓ and x_r respectively, and otherwise lie entirely outside of E. Thus, $x_j \notin E$. So, we see that $x_j \in E \iff y_j = 1$, and so $\mathbb{1}_E$ correctly classifies this labeling. Since this holds for all labelings of the n points, the VC dimension is $\geq n$. Since this holds for all $n \in \mathbb{N}$, we are done.
■

Problem 2

Consider classification problem using indicator functions of all half spaces in $\mathbb{R}^d$ (same as 1(a)):

$$\mathcal{F} := \{f(x) = \mathbb{1}_{\{w^\top x \leq t\}} : w \in \mathbb{R}^d,\, t \in \mathbb{R}\}.$$

Across this problem, we use 0-1 loss. Suppose that we indeed have the deterministic relation $Y = f^\star(X)$ holds under the true underlying data distribution D for certain $f^\star \in \mathcal{F}$. Prove that with probability at least $1-\delta$, the population risk of ERM is $\leq C\sqrt{\frac{d\log(n/\delta)}{n}}$, where C is some absolute constant and $n \geq 2$ is the number of training samples.

Solution

Proof. Let $r(\cdot)$ denote the population risk and $\hat{r}(\cdot)$ denote the empirical risk. We have from Lecture 4 that with probability $\geq 1-\delta$, the following bound holds:

$$\text{excess risk} \equiv r(\hat{f}) - r(f^*) \leq 4\mathcal{R}_n(\mathcal{F}) + 2B\sqrt{\frac{\log(1/\delta)}{2n}}$$

In the realizable setting, $r(f^*) = 0$, and so this is also a bound on the population risk of the ERM classifier $\hat{f}$. In the 0-1 classification setting, we know $B = 1$. Suppose first that $n > d$. Then, by Corollary 1 from Lecture 3, we may use our knowledge that the VC dimension of $\mathcal{F}$ is $d+1$ (see Problem 1(a)) to see

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log(2) + 2(d+1)\log(en/(d+1))}{n}}$$

Since $n > d$, we know that $2 < e \leq en/(d+1)$, and so $\log(2) \leq \log(en/(d+1))$. Thus,

$$\mathcal{R}_n(\mathcal{F}) \leq 2\sqrt{\frac{(d+1)\log(en/(d+1))}{n}}$$

We may certainly suppose that $\delta < \frac{d+1}{e}$, and so $\log(en/(d+1)) < \log(n/\delta)$. Since $d+1 \leq 2d$, we get

$$\mathcal{R}_n(\mathcal{F}) \leq 2\sqrt{2}\sqrt{\frac{d\log(n/\delta)}{n}}$$

So,

$$r(\hat{f}) \leq 2\sqrt{2}\sqrt{\frac{d\log(n/\delta)}{n}} + \frac{2}{\sqrt{2}}\sqrt{\frac{\log(1/\delta)}{n}}$$

Since $n \geq 1 \implies \log(1/\delta) \leq \log(n/\delta)$ and $d \geq 1$, we finally see

$$r(\hat{f}) \leq 2\sqrt{2}\sqrt{\frac{d\log(n/\delta)}{n}} + \sqrt{2}\sqrt{\frac{d\log(n/\delta)}{n}} = 3\sqrt{2}\sqrt{\frac{d\log(n/\delta)}{n}}$$

as desired. Suppose now that $2 \leq n \leq d$. By boundedness of the functions in $\mathcal{F}$, we see that the Rademacher complexity $\mathcal{R}_n(\mathcal{F}) \leq 1$ always (we have $\frac{1}{n}\left|\sum_i \epsilon_i f(x_i)\right| \leq \frac{1}{n}\sum_i |\epsilon_i f(x_i)| \leq \sup_x |f(x)| \leq 1$ since $|\epsilon_i| = 1$). So,

$$r(\hat{f}) \leq 4 + \sqrt{2}\sqrt{\frac{\log(1/\delta)}{n}} \leq 4 + \sqrt{2}\sqrt{\frac{d\log(n/\delta)}{n}}$$

In this setting, $1 \leq \frac{d}{n}$. We may certainly take $\delta < \frac{n}{e} \implies e < \frac{n}{\delta} \implies 1 < \log(n/\delta)$. Together, we see that $1 \leq \sqrt{\frac{d\log(n/\delta)}{n}}$, from which we get that

$$r(\hat{f}) \leq 4\sqrt{\frac{d\log(n/\delta)}{n}} + \sqrt{2}\sqrt{\frac{d\log(n/\delta)}{n}} = (4+\sqrt{2})\sqrt{\frac{d\log(n/\delta)}{n}}$$

Thus, if we take $C := \max\{4+\sqrt{2},\ 3\sqrt{2}\}$, the desired bound holds for all $n \geq 2$. ■

Problem 3

Consider $\mathcal{F}$ as the set of linear functions with bounded weights

$$\mathcal{F} = \left\{ f(x) = w^\top x : \, w \in \mathbb{R}^d, \, \|w\|_2 \leq 1 \right\}.$$

on domain $X = \left\{ x \in \mathbb{R}^d, \, \|x\|_2 \leq 1 \right\}$. Prove that there exists an absolute constant c, C s.t. for any $\epsilon \in (0,1)$:

$$\left(\frac{c}{\epsilon}\right)^d \leq N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \left(\frac{C}{\epsilon}\right)^d$$

Solution

Proof. We note that for any $f, g \in \mathcal{F}$ given by $w_f, w_g \in \mathbb{R}^d$ respectively, we have that

$$\|f - g\|_\infty = \sup_{\|x\|_2 \leq 1} |w_f^T x - w_g^T x| = \sup_{\|x\|_2 \leq 1} |(w_f - w_g)^T x|$$

This is maximized for $x^* = \frac{w_f - w_g}{\|w_f - w_g\|_2}$ obviously; in this case, we have $(w_f - w_g)^T x^* = \frac{\|w_f - w_g\|_2^2}{\|w_f - w_g\|_2}$, yielding

$$\|f - g\|_\infty = \|w_f - w_g\|_2$$

So, let $B := \{w \in \mathbb{R}^d : \ \|w\|_2 \leq 1\}$ denote the unit ball in $\mathbb{R}^d$. There is clearly a bijection between $\mathcal{F}$ and B, and the above logic reveals that this maps the $\|\cdot\|_\infty$ norm on $\mathcal{F}$ to the $\|\cdot\|_2$ norm on B. As such, $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) = N(\epsilon, B, \|\cdot\|_2)$.

We now seek to bound this covering number. Note that for any cover using n balls of radius ϵ, say with centers $\{x_1, \ldots, x_n\}$, we have $B \subseteq \bigcup_{j=1}^n B_\epsilon(x_j) \implies \operatorname{vol}(B) \leq \sum_{j=1}^n \operatorname{vol}(B_\epsilon(x_j)) = n\epsilon^d \operatorname{vol}(B)$, where we used that the volume of a union is upper bounded by the sum of the volumes (equality holds iff the union is almost disjoint). So, $n \geq \frac{1}{\epsilon^d}$ for all covers, and so this certainly holds for the minimal cover.

Now, we also know that the covering number is $\leq M(\epsilon, B, \|\cdot\|_2)$, the packing number, by Lemma 2 from Lecture 5. Consider any ϵ-packing of size n, which means we may fit n disjoint balls, say with centers $\{x_1, \ldots, x_n\} \subseteq B$, of radius $\frac{\epsilon}{2}$ inside B. Then, since these balls are disjoint and are contained in the closed ball of radius $1 + \frac{\epsilon}{2}$ about the origin (at worst case the center is on the boundary of B), we get

$$\bigsqcup_{j=1} B_{\epsilon/2}(x_j) \subseteq B_{1+\epsilon/2}(0) \implies \operatorname{vol}\left(\bigsqcup_{j=1} B_{\epsilon/2}(x_j)\right) = \sum_{j=1}^n \operatorname{vol}(B_{\epsilon/2}(x_j)) \leq \operatorname{vol}(B_{1+\epsilon/2}(0))$$

So, since $\operatorname{vol}(B_r(x)) = r^d \operatorname{vol}(B)$, this becomes

$$n\left(\frac{\epsilon}{2}\right)^d \leq \left(1 + \frac{\epsilon}{2}\right)^d \implies n \leq \left(\frac{2+\epsilon}{2} \cdot \frac{2}{\epsilon}\right)^d$$

For $\epsilon < 1$ we know $\frac{2+\epsilon}{2} \leq \frac{3}{2}$, and so

$$n \leq \left(\frac{3}{2} \cdot \frac{2}{\epsilon}\right) = \left(\frac{3}{\epsilon}\right)^d$$

Since this holds for every packing of size n, it also holds for the maximal packing. So, $M(\epsilon, B, \|\cdot\|_2) \leq \left(\frac{3}{\epsilon}\right)^d$. Thus,

$$\left(\frac{1}{\epsilon}\right)^d \leq N(\epsilon, B, \|\cdot\|_2) = N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \left(\frac{3}{\epsilon}\right)^d$$

and the result is proven. ■

Problem 4

Consider regression using function class $\mathcal{F}$, which is the set of all non-decreasing functions on domain $\mathbb{R}$ with range $Y = [-1, 1]$.

(a) Consider a fixed set of points $\{x_i\}_{i=1}^n$, and the corresponding distance:

$$\rho_n(f, g) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - g(x_i))^2}$$

Prove that for any $\epsilon \in (0, 1)$, we have $N(\epsilon, \mathcal{F}, \rho_n) \leq (n+1)^{\left(\frac{1}{\epsilon}+1\right)}$.

(b) Use (a) to prove the following bound on Rademacher complexity:

$$\mathcal{R}_n(\mathcal{F}) \leq c \cdot \sqrt{\frac{\log n}{n}}$$

for some absolute constant c.

(c) Use the above results to bound the excess risk of ERM with squared loss.

Solution

Proof. (a) Let $\epsilon \in (0, 1)$. We note that ρ_n measures the square root of average squared variation only on our test points. So, if we discretize the y-axis at scale $\frac{\epsilon}{2}$ (anything $< \epsilon$ suffices), functions taking the same discrete values at all of the x_j will have variation $< \epsilon$ at each x_j (less than $\frac{\epsilon}{2}$ in either direction), and so they will have ρ_n distance $< \epsilon$. Precisely, let $Y := \{-1, -1+2\epsilon, -1+4\epsilon, \ldots\} \subseteq [-1, 1]$; then, $|Y| = \lfloor \frac{2}{\epsilon} \rfloor$. We define a class of functions $\mathcal{G}_{n,\epsilon} \subseteq \mathcal{F}$ where for any set of n non-decreasing values $(y_1, \ldots, y_n)$ from Y, there is a function in $\mathcal{G}_{n,\epsilon}$ that realizes those values precisely at $x_1, \ldots x_n$ in the non-decreasing order (i.e. $g(x_j) = y_j \ \forall j$). It is easy to see that for any $f \in \mathcal{F}$, there is some non-decreasing sequence $\{y_1, \ldots, y_n\} \subseteq Y$ such that $|f(x_j) - y_j| < \epsilon$ for all j by construction of Y. Since there is some $g \in \mathcal{G}_{n,\epsilon}$ realizing this sequence of y_j's, we find that $|f(x_j) - g(x_j)| < \epsilon$ for all j, and so $\rho_n(f, g) < \epsilon$. Therefore, the set of ϵ-balls with centers in $\mathcal{G}_{n,\epsilon}$ covers $\mathcal{F}$. So, the minimal covering number must be less, yielding

$$N(\epsilon, \mathcal{F}, \rho_n) \leq |\mathcal{G}_{n,\epsilon}|$$

We now must bound the cardinality of $\mathcal{G}_{n,\epsilon}$. This cardinality is precisely equal to the number of non-decreasing sequences of length n that can be taken from Y. By stars and bars, this equals

$$|\mathcal{G}_{n,\epsilon}| = \binom{|Y|+n}{n} = \frac{(\lfloor \frac{2}{\epsilon} \rfloor + n)!}{n! \lfloor \frac{2}{\epsilon} \rfloor !} \leq \left(\frac{e(n + \lfloor \frac{2}{\epsilon} \rfloor)}{\lfloor \frac{2}{\epsilon} \rfloor}\right)^{\lfloor \frac{2}{\epsilon} \rfloor} \leq \left(\frac{\epsilon e n + 2e}{2}\right)^{\lfloor \frac{2}{\epsilon} \rfloor} = \left(\frac{e}{2}\right)^{\lfloor \frac{2}{\epsilon} \rfloor} \cdot (\epsilon n + 1)^{\lfloor \frac{2}{\epsilon} \rfloor}$$

I am unsure how to continue from here :)

(b) Consider a fixed set of points $\{x_i\}_{i=1}^n$. Using Theorem 2 from Lecture 5 and the bound from (a), we find

$$\mathcal{R}_n(\mathcal{F}(x_{1:n})) \leq \inf_{\alpha > 0}\left\{\alpha + \sqrt{\frac{2\log\left((n+1)^{(1+1/\alpha)}\right)}{n}}\right\} = \inf_{\alpha > 0}\left\{\alpha + \sqrt{\frac{2(1+1/\alpha)\log(n+1)}{n}}\right\}$$

We know that $n + 1 \leq n^2$ for all $n \in \mathbb{N}$, and so $\log(n+1) \leq 2\log(n)$. This gives

$$\mathcal{R}_n(\mathcal{F}(x_{1:n})) \leq \inf_{\alpha > 0}\left\{\alpha + 2\sqrt{1 + \frac{1}{\alpha}}\sqrt{\frac{\log(n)}{n}}\right\}$$

Problem 4 continued on next page. . .

Since $1 + \frac{1}{\alpha} \geq 1$, we know that $\sqrt{1 + \frac{1}{\alpha}} \leq 1 + \frac{1}{\alpha} = \frac{\alpha+1}{\alpha}$. Thus, letting $c_n := 2\sqrt{\frac{\log(n)}{n}}$,

$$\mathcal{R}_n(\mathcal{F}(x_{1:n})) \leq \inf_{\alpha>0} \left\{ \alpha + c_n \frac{\alpha+1}{\alpha} \right\} = \inf_{\alpha>0} \left\{ \frac{\alpha^2 + c_n\alpha + c_n}{\alpha} \right\}$$

The infimum is certainly less than or equal to the value at $\alpha = 1$. Plugging this in, we get

$$\mathcal{R}_n(\mathcal{F}(x_{1:n})) \leq \frac{1 + c_n + c_n}{1} = 2c_n = 4\sqrt{\frac{\log(n)}{n}}$$

Since this holds for all fixed sets of points, we get the result that

$$\mathcal{R}_n(\mathcal{F}) \leq 4\sqrt{\frac{\log(n)}{n}}$$

(c) Let $y, y', y^* \in Y$. Then, letting $\ell(\cdot, \cdot)$ be the squared loss,

$$|\ell(y, y^*) - \ell(y', y^*)| = |(y - y^*)^2 - (y' - y^*)^2| = |(y - y^* - y' + y^*)(y - y^* + y' - y^*)| = |y - y'| \cdot |y + y' - 2y^*|,$$

where we used the difference of squares. Since $|y + y' - 2y^*| \leq 4$ by boundedness of Y, we see that

$$|\ell(y, y^*) - \ell(y', y^*)| \leq 4|y - y'|$$

So, the square loss over this domain is 4-Lipschitz in the first slot. So, using Theorem 1 from Lecture 5, as well as part (b), we see that

$$\mathcal{R}_n(\ell \circ \mathcal{F}) \leq 16\sqrt{\frac{\log(n)}{n}}$$

Therefore, we get the bound

$$\text{excess risk} \leq 16\sqrt{\frac{\log(n)}{n}} + \text{small concentration terms}$$

■