CSCI 678: Theoretical Machine Learning Homework 2

Fall 2024, Instructor: Haipeng Luo

This homework is due on **10/13, 11:59pm**. *See course website for more instructions on finishing and submitting your homework as well as the late policy. Total points: 50*

1. (**Pseudo-dimension and fat-shattering dimension**) For a function $f : [0, 1] \to [-1, 1]$, define its total variation $V(f)$ as

$$V(f) = \sup_{\substack{1 \le m \in \mathbf{Z}_+ \\ 0 = x_0 < x_1 < \cdots < x_m = 1}} \sum_{j=1}^{m} |f(x_j) - f(x_{j-1})|,$$

which, intuitively, measures how much the function varies on the interval $[0, 1]$. Now, consider the function class $\mathcal{F} = \{f : [0, 1] \to [-1, 1] \mid V(f) \le B\}$ for some constant $B > 0$.

(a) (4pts) Prove that the Pseudo-dimension of $\mathcal{F}$ is infinity.

Proof. For any n, consider a sequence of n pairs $(x_1, y_1), \ldots, (x_n, y_n) \in [0, 1] \times [-1, 1]$ with $x_t = t/n$ and $y_t = 0$ for all t. Then, for any labeling $s_1, \ldots, s_n \in \{-1, +1\}$, consider the piece-wise constant function f (with at most n pieces) such that

$$f(x) = \frac{Bs_t}{2n}, \ \forall x \in \left(\frac{t-1}{n}, \frac{t}{n}\right],$$

and additionally $f(0) = f(1/n)$. Then, by construction, $\text{sign}(f(x_t) - y_t) = s_t$ trivially holds for all $t = 1, \ldots, n$. Moreover, the total variation of f is at most $\frac{B}{n} \times (n-1) < B$, since every two consecutive pieces of the function contribute $\frac{B}{n}$ variation. This shows $f \in \mathcal{F}$ and thus $\text{Pdim}(\mathcal{F}) = \infty$. $\square$

(b) Follow the two steps below to prove that the fat-shattering dimension of $\mathcal{F}$ at scale $\alpha \le 1$ is

$$\text{fat}(\mathcal{F}, \alpha) = 1 + \left\lfloor \frac{B}{\alpha} \right\rfloor.$$

i. (4pts) For $n \le 1 + \frac{B}{\alpha}$, construct a sequence of n pairs $(x_1, y_1), \ldots, (x_n, y_n) \in [0,1] \times [-1,1]$, such that for any labeling $s_1, \ldots, s_n \in \{-1, +1\}$, there exists $f \in \mathcal{F}$ with $s_t(f(x_t) - y_t) \ge \alpha/2$ for all $t = 1, \ldots, n$. (This shows $\text{fat}(\mathcal{F}, \alpha) \ge 1 + \lfloor \frac{B}{\alpha} \rfloor$.)

Proof. The construction is similar to the last question: consider $x_t = t/n$ and $y_t = 0$ for all t; for any labeling $s_1, \ldots, s_n \in \{-1, +1\}$, consider the piece-wise constant function f (with at most n pieces) such that

$$f(x) = \frac{\alpha s_t}{2}, \ \forall x \in \left(\frac{t-1}{n}, \frac{t}{n} \right],$$

and additionally $f(0) = f(1/n)$. Then, by construction we have for any t:

$$s_t(f(x_t) - y_t) = s_t f(x_t) = \frac{\alpha s_t^2}{2} = \frac{\alpha}{2}.$$

Moreover, the total variation of f is at most $\alpha \times (n-1) \le B$, since every two consecutive pieces of the function contribute α variation, which shows $f \in \mathcal{F}$. $\square$

ii. (5pts) For any $n > 1 + \frac{B}{\alpha}$ and any sequence of n pairs $(x_1, y_1), \ldots, (x_n, y_n) \in [0,1] \times [-1,1]$ with $x_1 < x_2 < \cdots < x_n$, show that if $f : [0,1] \to [-1,1]$ is such that $s_t(f(x_t) - y_t) \ge \alpha/2$ for all $t = 1, \ldots, n$ where

$$s_1 = -1, s_2 = +1, s_3 = -1, s_4 = +1, \ldots,$$

and $g : [0,1] \to [-1,1]$ is such that $s_t(g(x_t) - y_t) \ge \alpha/2$ for all $t = 1, \ldots, n$ where

$$s_1 = +1, s_2 = -1, s_3 = +1, s_4 = -1, \ldots,$$

then we must have $V(f) + V(g) > 2B$. (Convince yourself that this implies $\text{fat}(\mathcal{F}, \alpha) \le 1 + \lfloor \frac{B}{\alpha} \rfloor$.)

Proof. By the definition of total variation and the fact $x_1 < x_2 < \ldots < x_n$, we know $V(f) \ge \sum_{t=1}^{n-1} |f(x_t) - f(x_{t+1})|$. On the other than, by the stated condition, we have $f(x_t) \ge y_t + \alpha/2$ if t is even and $f(x_t) \le y_t - \alpha/2$ if t is odd. Therefore, for an odd t, we have

$$|f(x_t) - f(x_{t+1})| \ge f(x_{t+1}) - f(x_t) \ge y_{t+1} - y_t + \alpha,$$

and similarly, for an even t, we have

$$|f(x_t) - f(x_{t+1})| \ge f(x_t) - f(x_{t+1}) \ge y_t - y_{t+1} + \alpha.$$

On the other hand, by the same argument, $V(g)$ is at least $\sum_{t=1}^{n-1} |g(x_t) - g(x_{t+1})|$, and for an even t, we have

$$|g(x_t) - g(x_{t+1})| \ge g(x_{t+1}) - g(x_t) \ge y_{t+1} - y_t + \alpha,$$

and for an odd t, we have

$$|g(x_t) - g(x_{t+1})| \ge g(x_t) - g(x_{t+1}) \ge y_t - y_{t+1} + \alpha.$$

To sum up, for both even and odd t, the following holds:

$$|f(x_t) - f(x_{t+1})| + |g(x_t) - g(x_{t+1})| \ge 2\alpha,$$

and thus $V(f) + V(g) \ge 2(n-1)\alpha > 2B$ where the last step is due to $n > 1 + \frac{B}{\alpha}$. $\square$

2. (**Zero-covering number and shattering**) Consider a class of binary predictors $\mathcal{F} \subset \{-1, +1\}^{\mathcal{X}}$. The concept of zero-covering number $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})$ given an $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n is analogous to $|\mathcal{F}|_{x_{1:n}}|$, the cardinality of the projection of $\mathcal{F}$ on a dataset $x_{1:n}$ (in the statistical learning setting). However, there are some subtle differences between them. In particular, while $|\mathcal{F}|_{x_{1:n}}| = 2^n$ is equivalent to $x_{1:n}$ being shattered by $\mathcal{F}$, $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) = 2^n$ is *not* equivalent to $\boldsymbol{x}$ being shattered by $\mathcal{F}$. In this problem, you will explore why this is case. (Understanding what the questions below are asking you to do is already a good test to your understanding of the related concepts.)

   (a) (4pts) Prove that if $\mathcal{F}$ shatters $\boldsymbol{x}$, then we indeed have $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) = 2^n$. (Recall that $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) \le 2^n$ is always true, so this is really asking you to show $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) \ge 2^n$.)

   Proof. By the definition of shattering, for any path $\boldsymbol{\epsilon} \in \{-1, +1\}^n$, there exists a classifier, denoted by $f_{\boldsymbol{\epsilon}}$, such that $f_{\boldsymbol{\epsilon}}(\boldsymbol{x}_t(\boldsymbol{\epsilon})) = \epsilon_t$ for all t. Now, let V be a zero-cover of $\mathcal{F}|_{\boldsymbol{x}}$. Then for any two different paths $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}'$ (and the corresponding $f_{\boldsymbol{\epsilon}}$ and $f_{\boldsymbol{\epsilon}'}$), there exist $\boldsymbol{v} \in V$ and $\boldsymbol{v}' \in V$ such that on the corresponding path $f_{\boldsymbol{\epsilon}}$ agrees with $\boldsymbol{v}$ and $f_{\boldsymbol{\epsilon}'}$ agrees with $\boldsymbol{v}'$. The claim is that these two trees $\boldsymbol{v}$ and $\boldsymbol{v}'$ cannot be the same element of V. Indeed, let t be the first index such that $\epsilon_t \ne \epsilon'_t$. Then we have $\boldsymbol{v}_t(\boldsymbol{\epsilon}_{1:t-1}) = f(\boldsymbol{x}_t(\boldsymbol{\epsilon})) = \epsilon_t \ne \epsilon'_t = f(\boldsymbol{x}_t(\boldsymbol{\epsilon}')) = \boldsymbol{v}'_t(\boldsymbol{\epsilon}'_{1:t-1})$, but since $\boldsymbol{\epsilon}_{1:t-1}$ and $\boldsymbol{\epsilon}'_{1:t-1}$ are the same, we conclude that $\boldsymbol{v}$ and $\boldsymbol{v}'$ are two different trees. Therefore, for each different path $\boldsymbol{\epsilon}$, there is a corresponding different $\boldsymbol{v} \in V$, implying that $|V| \ge 2^n$. □

   (b) (4pts) Next, prove that $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) = 2^n$ does not necessarily mean that $\mathcal{F}$ shatters $\boldsymbol{x}$. Hint: consider a tree $\boldsymbol{x}$ with depth n being the VC-dimension of $\mathcal{F}$ and the leftmost path consisting of n points that are shattered by $\mathcal{F}$ (in the statistical learning sense).

   Proof. First, the tree $\boldsymbol{x}$ mentioned in the hint satisfies $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) = 2^n$. This is because $\mathcal{F}$ can realize all the 2^n possible labelings for the leftmost path by construction, so just to cover this path we already need 2^n different trees. However, there are many ways to construct the rest of $\boldsymbol{x}$ to make sure that it cannot be shattered by $\mathcal{F}$. For example, by simply setting the rightmost path of this tree to have one unique element, we cannot find an f to realize the labeling $(+1, +1, \cdots, +1, -1)$ for this path. This completes the proof. □

   (c) (4pts) Finally, prove that if $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) = 2^n$, then there must exist a tree $\boldsymbol{x}'$ of depth n that is shattered by $\mathcal{F}$. Hint: use Theorem 1 of Lecture 6, that is, the online analogue of Sauer's lemma. (Note that combining (a) and (c), we have

   $$\mathrm{Ldim}(\mathcal{F}) = \max\left\{ n : \max_{\boldsymbol{x} \text{ of depth } n} \mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) = 2^n \right\},$$

   which is analogous to $\mathrm{VCdim}(\mathcal{F}) = \max\{n : \max_{x_{1:n}} |\mathcal{F}|_{x_{1:n}}| = 2^n\}$.)

   Proof. Let d be the Littlestone dimension of $\mathcal{F}$. It suffices to prove $d \ge n$, because then by definition there must exist a tree $\boldsymbol{x}'$ of depth n that is shattered by $\mathcal{F}$. Indeed, if $d < n$, then we can use the fact $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) = 2^n$ together with the online analogue of Sauer's lemma to arrive at the following contradiction.

   $$2^n = \mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) \le \sum_{i=0}^{d} \binom{n}{i} < \sum_{i=0}^{n} \binom{n}{i} = 2^n.$$

   This finishes the proof. □

3. (**Littlestone dimension**) Consider $\mathcal{X} = \mathbb{R}^d$ and the class

$$\mathcal{F} = \left\{ f_{\theta,b}(x) = \begin{cases} +1, & \text{if } \langle \theta, x \rangle + b = 0 \\ -1, & \text{else} \end{cases} \;\middle|\; 0 \neq \theta \in \mathbb{R}^d, b \in \mathbb{R} \right\}.$$

which is a generalization of the simple class Eq. (5) in Lecture 5 from one dimension to general dimension. In words, it classifies all the points residing in the hyperplane $\langle \theta, x \rangle + b = 0$ as $+1$, and everything else as -1. Follow the steps below to show $\text{Ldim}(\mathcal{F}) = d$.

(a) (3pts) Construct a set of d points $x_1, \ldots, x_d \in \mathbb{R}^d$ that can be shattered by $\mathcal{F}$ (in the statistical learning sense), which shows $d \leq \text{VCdim}(\mathcal{F}) \leq \text{Ldim}(\mathcal{F})$.

Proof. Simply let the d points be the d standard basis vectors in $\mathbb{R}^d$: $e_1, \ldots, e_d$. Then for any $\epsilon_{1:n} \in \{-1, +1\}^n$, the parameters $\theta = (\epsilon_1, \ldots, \epsilon_n)$ and $b = -1$ satisfy for all $t = 1, \ldots, n$: $\langle \theta, x_t \rangle + b = \theta_t + b = \epsilon_t - 1$ and thus $f_{\theta,b}(x_t) = \epsilon_t$. This completes the proof. $\square$

(b) (4pts) For $d = 2$, show that no tree $\boldsymbol{x}$ of depth 3 can be shattered by $\mathcal{F}$. Hint: consider different cases for the three points on the rightmost path of $\boldsymbol{x}$: are they collinear (that is, on the same line)? are some of them identical?

Proof. For any tree of depth 3, consider the three points on its rightmost path. If they are not collinear, then $\boldsymbol{\epsilon} = (+1, +1, +1)$ can not be realized by any classifiers in $\mathcal{F}$ since a line in $\mathbb{R}^2$ cannot pass through 3 points that are not collinear.

If they are collinear, there are two cases. First, if the first two points are identical, then $\boldsymbol{\epsilon} = (+1, -1, ?)$ (the value of ? does not matter) cannot be realized by any $f \in \mathcal{F}$ since it requires labeling the same point by $+1$ and -1 simultaneously.

On the other hand, if the first two points are distinct, then $\boldsymbol{\epsilon} = (+1, +1, -1)$ cannot be realized since the third point must be on the line that passes through the first two points, which means that any $f \in \mathcal{F}$ that labels the first two points as $+1$ must label the last point as $+1$ as well. To sum up, no tree of depth 3 can be shattered by $\mathcal{F}$. $\square$

(c) (8pts) Generalize the idea from the last question to show that for any dimension d, no tree of depth $d + 1$ can be shattered by $\mathcal{F}$, which shows $\text{Ldim}(\mathcal{F}) \leq d$. Hint: a set of n points $x_1, \ldots, x_n \in \mathbb{R}^d$ are *affinely* dependent if the following $n - 1$ points are linearly dependent: $x_1 - x_n, x_2 - x_n, \ldots, x_{n-1} - x_n$; convince yourself that two points being affinely dependent if and only if they are identical, and three points being affinely dependent if and only if they are collinear.

Proof. For any tree of depth $d + 1$, let $x_1, \ldots, x_{d+1}$ be the points on its rightmost path. If they are affinely independent, then no hyperplane can pass through all of them and thus no $f \in \mathcal{F}$ can realize $\boldsymbol{\epsilon} = (+1, +1, \ldots, +1)$. More formally, suppose that $f_{\theta,b}$ predicts $+1$ on all these points, that is, $\langle \theta, x_t \rangle + b = 0$ for all t. Then, we have $\langle \theta, x_t - x_{d+1} \rangle = 0$ for all t. Since the space $\{x \in \mathbb{R}^d : \langle \theta, x \rangle = 0\}$ is $(d - 1)$-dimensional, the d points $x_1 - x_{d+1}, x_2 - x_{d+1}, \ldots, x_d - x_{d+1}$ must be linearly dependent. This is a contraction to $x_{1:d+1}$ being affinely independent.

Now suppose $x_{1:d+1}$ are affinely dependent. In particular, let $k \geq 1$ be the smallest index such that $x_{1:k+1}$ are affinely dependent. Then we claim that no $f \in \mathcal{F}$ can realize $\boldsymbol{\epsilon} = (+1, +1, \ldots, -1, ?, \cdots, ?)$ where the first -1 appears on the $(k + 1)$-th coordinate (the value of ? does not matter), that is , no $f \in \mathcal{F}$ can predict $+1$ on $x_1, \ldots, x_k$ while predicting -1 on x_{k+1}. Indeed, suppose that θ and b are such that $\langle \theta, x_t \rangle + b = 0$ for $t = 1, \ldots, k$. Since $x_{1:k+1}$ are affinely dependent, there exist coefficients $a_1, \ldots a_k \in \mathbb{R}$, not all zero, such that $\sum_{t=1}^k a_t(x_t - x_{k+1}) = 0$, or equivalently, $\sum_{t=1}^k a_t x_t = (\sum_{t=1}^k a_t) x_{k+1}$. Multiplying both sides by θ and adding $(\sum_{t=1}^k a_t) b$ to both sides shows $(\sum_{t=1}^k a_t)(\langle \theta, x_{k+1} \rangle + b) = 0$. Since $\sum_{t=1}^k a_t \neq 0$ (otherwise $x_{1:k}$ are affinely dependent already, contradicting with the definition of k), we have $\langle \theta, x_{k+1} \rangle + b = 0$, which shows that no $f \in \mathcal{F}$ can predict $+1$ on $x_1, \ldots, x_k$ while predicting -1 on x_{k+1}. $\square$

4. (**Lower bound for online classification**) In this exercise you will prove $\mathcal{V}^{\text{seq}}(\mathcal{F}, n) \geq \sqrt{\frac{d}{8n}}$ where $d = \text{Ldim}(\mathcal{F}) \leq n$. For simplicity, we will further assume that n is a multiple of d. The construction of the environment is as follows. The labels $y_1, \ldots, y_n$ are i.i.d. Rademacher random variables. To define the example $x_1, \ldots, x_n$, we divide the entire n rounds evenly into d epochs, where epoch k contains rounds $n(k-1)/d + 1, \ldots, nk/d$. On the same epoch, x_t stays the same. Specifically, let $\epsilon_k = \text{sign}\left(\sum_{t \in \text{epoch } k} y_t\right)$ be the majority vote of the true labels in epoch k, that is,

$$\epsilon_k = \begin{cases} +1, & \text{if } \sum_{t \in \text{epoch } k} y_t \geq 0, \\ -1, & \text{else,} \end{cases}$$

and $\boldsymbol{x}$ be a tree of depth d that is shattered by $\mathcal{F}$. Then $x_t = \boldsymbol{x}_k(\boldsymbol{\epsilon})$ for any t that belongs to epoch k. This concludes the construction of the environment.

(a) (2pts) For any online learner, let $s_1, \ldots, s_n \in \{-1, +1\}$ be its sequential predictions for $x_1, \ldots, x_n$ in this environment. Calculate the learner's expected loss $\mathbb{E}\left[\sum_{t=1}^n \mathbf{1}\{s_t \neq y_t\}\right]$, where the expectation is with respect to the randomness of both the learner and the environment.

Proof. The answer is clearly $n/2$ since each y_t is an i.i.d. Rademacher random variable and s_t is independent of y_t. □

(b) (4pts) Calculate $\mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^n \mathbf{1}\{f(x_t) \neq y_t\}\right]$, the expected loss of the best classifier in $\mathcal{F}$, where the randomness is with respect to the randomness of the environment.

Proof. By the construction and the definition of shattering, there exists $f \in \mathcal{F}$ such that it correctly predicts all the majority votes $\epsilon_1, \ldots, \epsilon_d$, which also implies that it must be the best classifier. On epoch k, the number of mistakes this optimal classifier makes is the size of the minority, which is precisely $\frac{\frac{n}{d} - |\sum_{t \in \text{epoch } k} y_t|}{2}$. Summing over d epochs shows

$$\mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^n \mathbf{1}\{f(x_t) \neq y_t\}\right] = \frac{n}{2} - \frac{\mathbb{E}\left[\sum_{k=1}^d |\sum_{t \in \text{epoch } k} y_t|\right]}{2}.$$

□

(c) (4pts) Conclude the statement $\mathcal{V}^{\text{seq}}(\mathcal{F}, n) \geq \sqrt{\frac{d}{8n}}$. Hint: use the Khinchine inequality that says the expected magnitude of the sum of m i.i.d. Rademacher random variables is at least $\sqrt{m/2}$ for any $m \geq 1$.

Proof. Direct calculation shows

$$\begin{aligned} \text{Reg}(\mathcal{F}, n) &= \mathbb{E}\left[\sum_{t=1}^n \mathbf{1}\{s_t \neq y_t\}\right] - \mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^n \mathbf{1}\{f(x_t) \neq y_t\}\right] \\ &= \frac{\mathbb{E}\left[\sum_{k=1}^d |\sum_{t \in \text{epoch } k} y_t|\right]}{2} \geq \frac{d \cdot \sqrt{\frac{n}{2d}}}{2} = \sqrt{dn/8}, \end{aligned}$$

where the inequality is by the Khinchine inequality. Since this holds for any learner, normalizing proves $\mathcal{V}^{\text{seq}}(\mathcal{F}, n) \geq \sqrt{\frac{d}{8n}}$. □