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.

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

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

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

ii. (5pts) For any $n > 1 + \frac{B}{\alpha}$ and any sequence of n pairs $(x_1, y_1), \dots, (x_n, y_n) \in [0,1] \times [-1,1]$ 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 $t = 1, \dots, n$ where

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

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

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

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

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 $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) \leq 2^n$ is always true, so this is really asking you to show $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) \geq 2^n$.)

   (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).

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

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 $d \leq \text{VCdim}(\mathcal{F}) \leq \text{Ldim}(\mathcal{F})$.

(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?

(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.

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.

(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.

(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 $m \geq 1$.