Theoretical Machine Learning
Homework 2

Instructor: Haipeng Luo

This homework is due on **10/20, 11:59pm**. See course website for more instructions on finishing and submitting your homework as well as late policy.

1. (**Littlestone dimension**)

   (a) (6pts) Prove that if a class $\mathcal{F} \subset \{-1,+1\}^{\mathcal{X}}$ shatters an $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n, then the zero-covering number $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}})$ is exactly 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) (6pts) Is the converse of the last statement true? That is, does $\mathcal{N}_0(\mathcal{F}|_{\boldsymbol{x}}) = 2^n$ always imply that $\mathcal{F}$ shatters $\boldsymbol{x}$ (n is again the depth of $\boldsymbol{x}$)? Prove or disprove it.

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

   (a) (2pts) For any learner, let $s_1, \ldots, s_n \in \{-1,+1\}$ be its predictions for $x_1, \ldots, x_n$. Calculate the learner's expected loss $\mathbb{E}\left[\sum_{t=1}^n \mathbf{1}\{s_t \neq y_t\}\right]$.

   (b) (5pts) 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.

   (c) (3pts) Conclude the statement $\mathcal{V}^{\text{seq}}(\mathcal{F}, n) \geq \sqrt{\frac{d}{8n}}$ by using Khinchine's inequality which says $\mathbb{E}\left[\left|\sum_{t \in \text{epoch } k} y_t\right|\right] \geq \sqrt{\frac{n}{2d}}$.

3. (**Halving and Hedge**) (5pts) For a finite class of binary classifier $\mathcal{F}$, under the realizable assumption $\inf_{f \in \mathcal{F}} \sum_{t=1}^n \mathbf{1}\{f(x_t) \neq y_t\} = 0$, prove that Hedge with learning rate $\eta = 1/2$ makes at most $4 \ln |\mathcal{F}|$ mistakes in expectation, similar to the guarantee of Halving. (Hint: use Lemma 1 of Lecture 6.)

4. (**Learning classes with finite Littlestone dimension**) In this exercise you will complete the details of constructing a general algorithm that learns any class with regret $\mathcal{O}\left(\sqrt{dn \ln n}\right)$ where

$d = \text{Ldim}(\mathcal{F})$. As mentioned, we will "simulate" the cover of $\mathcal{F}$ for the unknown tree chosen by the environment using variants of the generalized Halving. Specifically, for each $m = 0, 1, \ldots, d$ and a sequence of m time steps $1 \le t_1 < t_2 < \cdots < t_m \le n$, we define an expert as follows:

Parameters: a set $\mathcal{T}$ of m time steps $1 \le t_1 < t_2 < \cdots < t_m \le n$ for some $m \le d$.
Let $\mathcal{F}' = \mathcal{F}$. For $t = 1, \ldots, n$,

- receive x_t and define

$$\mathcal{F}'_- = \{f \in \mathcal{F}' \mid f(x_t) = -1\} \quad \text{and} \quad \mathcal{F}'_+ = \{f \in \mathcal{F}' \mid f(x_t) = +1\}.$$

- predict $\operatorname{argmax}_{y \in \{-,+\}} \text{Ldim}(\mathcal{F}'_y)$ if $t \notin \mathcal{T}$, otherwise predict the opposite label; let $y_t^{\mathcal{T}}$ be this prediction.
- update $\mathcal{F}' \leftarrow \mathcal{F}'_{y_t^{\mathcal{T}}}$.

Let V be the set of all possible experts of this kind. The final algorithm is to run Hedge over this set of experts, that is, at time t, sample an expert with parameter $\mathcal{T}$ with probability proportional to $\exp\left(-\eta \sum_{\tau=1}^{t-1} \mathbf{1}\left\{y_\tau^{\mathcal{T}} \neq y_\tau\right\}\right)$, and then follow the prediction of this expert.

(a) (6pts) Prove that for each $f \in \mathcal{F}$, there exists an expert with some parameter $\mathcal{T}$ such that their predictions exactly match, that is, $y_t^{\mathcal{T}} = f(x_t)$ for all $t = 1, \ldots, n$. (Hint: consider what happens when we run the generalized Halving (Figure 1 of Lecture 6) on the sequence $(x_1, f(x_1)), \ldots, (x_n, f(x_n))$).

(b) (3pts) Calculate the size of the expert pool V.

(c) (4pts) Using the Hedge guarantee, conclude that this algorithm ensures

$$\mathbb{E}[\text{Reg}(\mathcal{F}, n)] = \mathcal{O}\left(\sqrt{dn \ln n}\right).$$