CSCI 678: Theoretical Machine Learning
Homework 3

Fall 2024, Instructor: Haipeng Luo

This homework is due on **11/03, 11:59pm**. *See course website for more instructions on finishing and submitting your homework as well as the late policy. Total points: 40*

1. (**Hedge**) (6pts) For a finite class of binary classifier $\mathcal{F} \subset \{-1,+1\}^{\mathcal{X}}$, 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. Hint: use Lemma 1 of Lecture 6. (Note that this is similar to the guarantee of Halving, but achieved via a proper algorithm this time.)

2. (**Perceptron and sequential fat-shattering dimension**) Recall the sequential fat-shattering dimension $\mathrm{sfat}(\mathcal{F}, \alpha)$ defined in Lectures 6. Let $\mathcal{X} = B_2^d$ and $\mathcal{F} = \left\{ f_\theta(x) = \langle \theta, x \rangle \mid \theta \in B_2^d \right\}$. In this exercise, you will prove $\mathrm{sfat}(\mathcal{F}, \alpha) \leq \frac{16}{\alpha^2}$ (which is independent of d) for any $\alpha > 0$, using an indirect approach that leverages the guarantee of the Perceptron algorithm.

   More specifically, suppose that $\boldsymbol{x}$ is a $\mathcal{X}$-valued tree of depth n that is α-shattered by $\mathcal{F}$, with witness $\boldsymbol{y}$, a $[-1, +1]$-valued tree. Now, imagine running Perceptron in the following problem instance in $\mathbb{R}^{d+1}$:

   Let $\theta' = \mathbf{0} \in \mathbb{R}^{d+1}$. For $t = 1, \ldots, n$:
   - Environment reveals example $x'_t = \frac{1}{\sqrt{2}}(\boldsymbol{x}_t(y'_{1:t-1}), \boldsymbol{y}_t(y'_{1:t-1})) \in B_2^{d+1}$.
   - Perceptron algorithm predicts $s_t = \mathrm{sign}(\langle x'_t, \theta' \rangle)$.
   - Environment reveals $y'_t = -s_t$, forcing Perceptron to make an update $\theta' \leftarrow \theta' + y'_t x'_t$.

   Note that the environment is valid even though it seemingly decides the label y'_t after seeing the algorithm's prediction s_t, since Perceptron is a deterministic algorithm (and thus $x'_{1:n}$ and $y'_{1:n}$ are in fact all fixed ahead of time).

   (a) (4pts) Prove that the data constructed above satisfy the γ-margin assumption (Assumption 1 of Lecture 7) with $p = q = 2$. In other words, find a specific value of $\gamma > 0$ and show that there exists $\theta'_\star \in B_2^{d+1}$ such that $y'_t \langle \theta'_\star, x'_t \rangle \geq \gamma$ holds for all $t = 1, \ldots, n$.

   (b) (3pts) Use the guarantee of Perceptron (that is, Theorem 3 of Lecture 7) to conclude $\mathrm{sfat}(\mathcal{F}, \alpha) \leq \frac{16}{\alpha^2}$.

3. (**Winnow**) When the γ-margin assumption holds with $p = q = 2$, we have seen that Perceptron makes at most $\frac{1}{\gamma^2}$ mistakes for an online binary classification problem. In this exercise, you will prove a similar result when the γ-margin assumption holds with $p = 1$ and $q = \infty$, using a different algorithm called *Winnow*. To show this, we first consider the following generalization of Perceptron, defined in terms of some *link function* $g : \mathbb{R}^d \to \mathbb{R}^d$.

Algorithm 1: A generalization of Perceptron

Let $\theta = \mathbf{0}$. For $t = 1, \ldots, n$:

- Receive x_t and predict $s_t = \text{sign}(\langle x_t, g(\theta)\rangle)$.
- Receive $y_t \in \{-1, +1\}$. If $y_t \neq s_t$, update $\theta \leftarrow \theta + y_t x_t$.

It is clear that when instantiated with g being the identity mapping $g(\theta) = \theta$, Algorithm 1 is exactly the Perceptron algorithm. Below, we will see that the Winnow algorithm is also an instance of Algorithm 1 but with a different link function. Throughout, we assume $x_t \in B_\infty^d$, that is, $\|x_t\|_\infty \leq 1$, for all t.

(a) Consider running Algorithm 1 with link function $g(\theta) = \exp(\eta\theta)$ and some parameter $\eta > 0$ (where the exponentiation is applied coordinate-wise to the vector $\eta\theta$). Let's call this the simplified Winnow algorithm.

i. (4pts) Find a sequence of loss vectors $\ell_1, \ldots, \ell_n \in [-1, +1]^d$ such that the prediction of simplified Winnow $s_t = \text{sign}(\langle x_t, g(\theta)\rangle)$ can be equivalently written as $s_t = \text{sign}(\langle x_t, p_t\rangle)$, where $p_t \in \Delta(d)$ is a distribution such that

$$p_t(i) \propto \exp\left(-\eta \sum_{\tau < t} \ell_\tau(i)\right), \quad \text{for all } i = 1, \ldots, d.$$

ii. (8pts) Based on the reformulation of the last question, apply Lemma 1 of Lecture 6 to show that as long as $\eta \leq 1$, we have for any $\theta^\star \in \Delta(d)$:

$$\sum_{t=1}^{n} \mathbf{1}\{y_t \neq s_t\} y_t \langle \theta^\star, x_t\rangle \leq \frac{\ln d}{\eta} + \eta M,$$

where $M = \sum_{t=1}^{n} \mathbf{1}\{y_t \neq s_t\}$ is the total number of mistakes made by the simplified Winnow algorithm.

iii. (3pts) Consider the following assumption that is slightly stronger than the original γ-margin assumption with $p = 1$ and $q = \infty$:

$$\text{there exists } \theta^\star \in \Delta(d) \text{ such that } y_t \langle \theta^\star, x_t\rangle \geq \gamma \text{ for all } t. \tag{1}$$

Prove that under this assumption, the total number of mistakes M made by the simplified Winnow algorithm is at most $\frac{4 \ln d}{\gamma^2}$ when $\eta = \frac{\gamma}{2} \leq 1$.

(b) Now consider the original γ-margin assumption, that is:

$$\text{there exists } \theta^\star \in B_1^d \text{ such that } y_t \langle \theta^\star, x_t\rangle \geq \gamma \text{ for all } t. \tag{2}$$

To deal with this more general case, we will run Algorithm 1 using a different link function $g(\theta) = \exp(\eta\theta) - \exp(-\eta\theta)$ (again, the exponentiation is coordinate-wise). This is the (actual) Winnow algorithm.

i. (4pts) Prove that the Winnow algorithm is the same as running the simplified Winnow algorithm over examples $x'_t = (x_t, -x_t) \in B_\infty^{2d}$ and $y'_t = y_t$ for $t = 1, \ldots, n$.

ii. (6pts) Under the margin assumption Equation (2), further prove that the examples $(x'_{1:n}, y'_{1:n})$ defined above satisfy Equation (1) for some margin γ', that is, there exists $\theta' \in \Delta(2d)$ such that $y'_t \langle \theta', x_t\rangle \geq \gamma'$ for all t.

iii. (2pts) Finally, under the margin assumption Equation (2), use the result from Question (a)iii to provide a bound on the total number of mistakes made by the Winnow algorithm when $\eta = \frac{\gamma}{2}$.