CSCI699: Theory of Machine Learning Fall 2021

Lecture 17: Online Learning

Instructor: Vatsal Sharan *Scribe: Sophie Hsu*

Recap:

Compared to PAC learning, online learning relaxes the distribution by making no statistical assumptions regarding the origin of the sequence of examples.

At every time step t

- $\rightarrow$ learner receives an input $x_t \in X$.
- $\rightarrow$ makes prediction $p_t \in Y$
- $\rightarrow$ see new label $y_t \in Y$, suffer loss $\ell(p_t, y_t)$.

Think $y = \{0, 1\}$, $\ell(p_t, y_t) = \mathbb{1}\{p_t \neq y_t\}$.

1 Realizability

Definition 1. *Mistake bound model: Let $\mathcal{H}$ be a hypothesis class and A be a online learning algorithm. Given any sequence $S = (x_1, h^*(x_1)), \ldots, (x_T, h^*(x_T))$ of T labelled datapoints where $h^* \in \mathcal{H}$, let $M_A(S)$ be the number of mistakes A makes on the sequence S. We denote by $M_A(\mathcal{H})$ to the supremum of $M_A(S)$ over all possible S.*

If there exists an algorithm A that satisfies a mistake bound of the form $M_A(\mathcal{H}) \leq B < \infty$, we say $\mathcal{H}$ is online learnable in the mistake bound model.

Algorithm 1 Consistent

Initialize $V_1 = \mathcal{H}$
for for $t = 1, \ldots, T$ **do**
  receive x_t
  choose any $h \in V_t$
  predict $p_t = h(x_t)$
  receive $y_t = h^*(x_t)$, loss $\mathbb{1}(p_t \neq y_t)$
  update $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$

The consistent algorithm will remove at least one hypothesis from V_t when there is a mistake prediction. Therefore, after M mistakes, $|V_t| \leq |\mathcal{H}| - M$ and would be ≤ 1 since V_t is nonempty. Hence,

Proposition 2. *Let $\mathcal{H}$ be a finite hypothesis class. The above algorithm gets a mistake bound*

$$M_{consistent}(\mathcal{H}) \leq |\mathcal{H}| - 1$$

The following is a smarter algorithm to choose $h \in V_t$ with fewer mistakes.

Algorithm 2 Halving

Initialize $V_1 = \mathcal{H}$
for for $t = 1, \ldots, T$ **do**
receive x_t
predict $p_t = \text{argmax}_{r \in \{0,1\}} |\{h \in V_t : h(x_t) = r\}|$ (if tie, $p_t = 1$)
receive $y_t = h^*(x_t)$, loss $\mathbb{1}(p_t \neq y_t)$
update $V_{t+1} = \{h \in v_t : h(x_t) = y_t\}$.

Proposition 3. *Let $\mathcal{H}$ be a finite hypothesis class. Then halving algorithm satisfies the mistake bound $M_{Halving}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$*

Proof. Whenever the algorithm errs, we have $|V_{t+1}| \leq \frac{|V_t|}{2}$. If M is number of mistakes,

$$|V_{T+1}| \leq |\mathcal{H}| 2^{-M}$$

as $\quad |V_{T+1}| \geq 1$

$$\Rightarrow M \leq \log_2(\mathcal{H})$$

□

Next, we aim to characterize online learning. Nick Littlestone presented a dimension of hypothesis classes that characterizes the best achievable mistake bound.

Littlestone dimension

Idea: view online learning as a 2-player game between learner and the environment.

At time t :

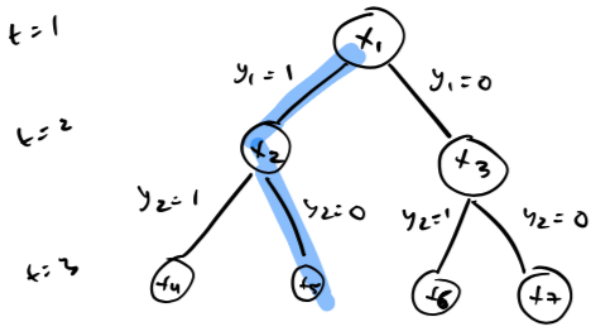
→ environment picks x_t

→ learner picks p_t

→ environment picks $y_t = 1 - p_t$

Q) How to choose x_t to get the learner to make maximum number of mistakes, while ensuring realizability?

Strategy for the environment can be formally described as a binary tree. Each node of the tree is associated with an instance from X. If the learner predicts $p_t = 1$ the environment will declare that this is a wrong prediction (i.e., $y_t = 0$) and will traverse to the right child of the current node. If the learner predicts pt $= 0$ then the environment will set $y_t = 1$ and will traverse to the left child. This process will continue and at each round.



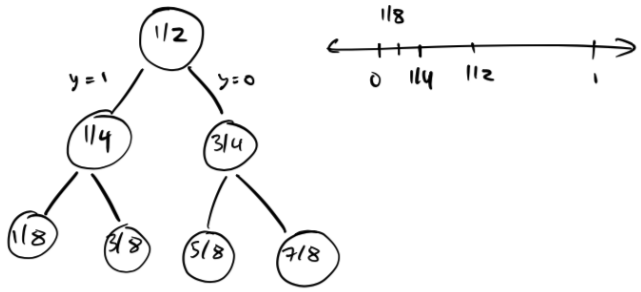
Definition 4. *$\mathcal{H}$ shattered tree: A shattered tree of depth d is a sequence of instances $x_1, x_2, \ldots, x_{2^d-1} \in X$ such that for every path from the root to a leaf, $\exists\ h \in \mathcal{H}$ which realizes all the labels along this path.*

Definition 5. *Littestone dimension: Littlestone dimension of hypothesis class $\mathcal{H}(Ldim(\mathcal{H}))$ is the maximal T such that $\exists$ a tree of depth T shattered by $\mathcal{H}$.*

Lemma 6. *For any online learning algorithm A, $M_A(\mathcal{H}) \geq Ldim(\mathcal{H})$.*

Examples

1. Let $\mathcal{H}$ be a finite hypothesis class. Then $Ldim(\mathcal{H}) \leq \log_2(\mathcal{H})$.
   $\rightarrow$ Any tree which is shattered by $\mathcal{H}$ must have $|\mathcal{H}| \geq \#$ leaves in the tree.
2. $x = [0, 1]$, $\mathcal{H} = \{x \mapsto \mathbb{1}(x < a), a \in [0, 1]\}$ (thresholds on $[0, 1]$). Then, $Ldim(\mathcal{H}) = \infty$



As a direct corollary of the above bound, we have that thresholds are not online learnable.

Corollary 7. *Cannot learn thresholds in the online learning model.*

This follows because algorithm can have a mistake bound smaller than $Ldim(H)$. Next we see that there is in fact an algorithm which matches the Littlestone dimension bound.

Lemma 8. *There exists an algorithm A with $M_A(\mathcal{H}) \leq Ldim(\mathcal{H})$.*

Instead of predicting according to the larger class as done in Halving, we present an algorithm that predicts according to the class with larger Ldim.

Algorithm 3 Standard optimal algorithm (SOA)

Initialize $V_1 = M$
for for $t = 1, \ldots, T$ **do**
  receive x_t
  for $r \in \{0, 1\}$, let $V_t^{(r)} = \{h \in V_t : h(x_t) = r\}$
  predict $p_t = \text{argmax}_{r \in \{0,1\}} Ldim(v_t^{(r)})$
  receive $y_t = h^*(x_t)$, loss $\mathbb{1}(p_t \neq y_t)$
  update $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$.

Claim 9. *Whenever algorithm makes a mistake* $Ldim(V_{t+1}) = Ldim(V_t) - 1$

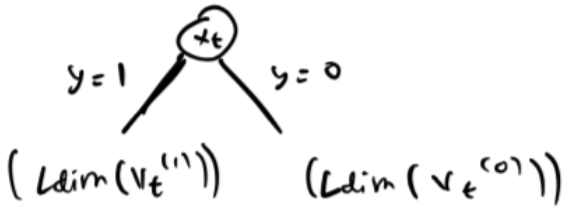
Proof. Proof by contradiction For the sake of contradiction, assume $Ldim(V_{t+1}) = Ldim(V_t)$ though the algorithm has made mistake. WLOG assume that the algorithm predicts 0. Then $y_t = 1$ since algorithm made a mistake.

$$Ldim(v_{t+1}) = L_{dim}(V_t^{(1)}) = Ldim(V_t)$$

also

$$\begin{aligned} Ldim(V_t^{(0)}) \geq & Ldim(V_t^{(1)}) = Ldim(V_t) \\ & Ldim(V_t^{(0)}) \leq Ldim(V_t) \\ \therefore \quad & Ldim(V_t^{(0)}) = Ldim(V_t) \end{aligned}$$

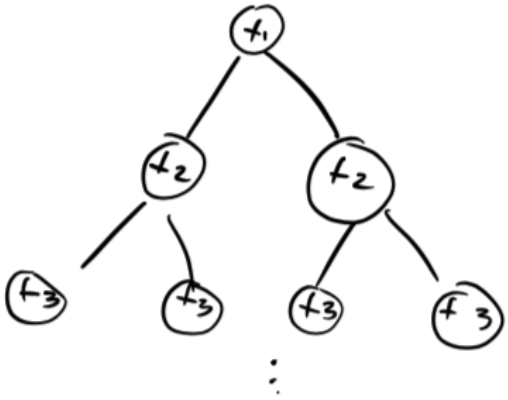
But then $Ldim(V_t) = Ldim(V_t^{(0)}) + 1$:



□

Theorem 10. *(Comparison to VCdim(*$\mathcal{H}$*)) For any class* $\mathcal{H}$*,* $VCdim(\mathcal{H}) \leq Ldim(\mathcal{H})$*. Further, the gap can be arbitrarily large.*

Proof. Suppose $\text{VCdim}(\mathcal{H}) = d$. Let $x_1, \ldots, x_d$ be shattered set



The gap can be arbitrarily large just because of thresholds. □

Aside: Some recent work shows that every $\mathcal{H}$ with finite $Ldim(\mathcal{H})$ can be learned privately (and vice versa) [1].

2 Online learning in the unrealizable case

Definition 11. *Regret: The regret of an algorithm A relative to a hypothesis h when run on a sequence of T examples is:*

$$Regret_A(h, T) = \sup_{(x_1,y_1),\cdots,(x_T,y_T)} \left[\sum_{t=1}^{T} |p_t - y_t| - \sum_{t=1}^{T} |h(x_t) - y_t| \right].$$

The regret of A relative to a hypothesis class $\mathcal{H}$ is:

$$Regret_A(\mathcal{H}, T) = \sup_{h \in \mathcal{H}} Regret_A(h, T).$$

*Note that if the sequence is realizable, this is the same as the mistake bound.

Q) Can we get sublinear regret ($o(T)$)? Unfortunately, no.

Consider $\mathcal{H} = \{h_0, h_1\}$, h_0 always predicts 0 and h_1 always predicts 1. An adversary can force the number of mistakes made by the algorithm to T.

But, the best predictor in hindsight is the majority of $y_1 \dots y_T$ which makes $\leq T/2$ mistakes.
$\Rightarrow$ Regret $\geq T/2$

To get around this, we allow randomized algorithms. Importantly, the environment decides y_t before the random coins are flipped.

Setup At every time step f,

$\rightarrow$ learner receives $x_t \in X$

$\rightarrow$ learner decides $p_t \in [0,1]$, probability of label being 1

$\rightarrow$ environment "decides" true label $y_t \in \{0,1\}$

$\rightarrow$ learner outputs $\hat{y}_t = \begin{cases} 1, & \text{w.p.} \quad p_t \\ 0, & \text{w.p.} \quad 1-p_t \end{cases}$

$\rightarrow$ Expected loss at time t,

$$\begin{aligned} \mathbb{P}\left(\hat{y}_t \neq y_t\right) &= \begin{cases} p_t, & \text{if} \quad y_t = 0 \\ 1-p_t & \text{, if} \quad y_t = 1 \end{cases} \\ &= |p_t - y_t| \end{aligned}$$

$$Regret_A(\mathcal{H}, T) = \sup_{h \in \mathcal{H}} \sup_{(x_1,y_1),,(x_T,y_T)} \left[\sum_{t=1}^{T} |p_t - y_t| - \sum_{t=1}^{T} |h(x_t) - y_t| \right]$$

Q) Can we get sublinear regret? Yes, using the *Weighted-Majority* algorithm.

Setting: Prediction with "expert advice". At every time t, learner has to choose one among d experts to predict based on. We then see true label, and the loss each expert has on that time step, which can use to retrieve future predictions.

Q) How well can we do compared to the best expert in hindsight?

Algorithm 4 Weighted Majority (also known as Multiplicative Weights/Hedge)

initialize $w^{(1)} = (1, \ldots, 1)$ (d dimensional)
for for $t = 1 \ldots$ T **do**
  set $\tilde{w}^{(t)} = w^{(t)} \mid z_t$ where $z_t = \Sigma_i w_i^{(t)}$
  choose expert i at random according to $P[i] = \tilde{w}_i^{(t)}$
  receive costs of all expects $v_t \in [0,1]^d$
  pay expected cost: $< \tilde{w}^{(t)}, v_t >$
  update: $\forall i : w_i^{(t+1)} = w_i^{(t)} e^{-\eta v_{t,i}}$

Theorem 12. *Assuming $T > 2\log(d)$, the Weighted-Majority algorithm enjoys the bound*

$$\sum_{t=1}^{T} < w^{(t)}, v_t > - \min_{i \in [d]} \sum_{t=1}^{T} v_{t,i} \leq \sqrt{2\log(d)T}$$

Proof. We have:

$$\begin{aligned} \log\left(\frac{Z_{t+1}}{Z_t}\right) &= \log\left(\frac{\sum_i \tilde{w}_i^{(t)} e^{-\eta v_{t,i}}}{Z_t}\right) \\ &= \log\left(\sum_i \hat{w}_i^{(t)} e^{-\eta v_{t,i}}\right) \end{aligned}$$

Using

1. $e^{-a} \leq 1 - a + \frac{a^2}{2}, \forall a \in (0,1)$

2. $\sum_i w_i^t = 1$

3. $\log(1-b) \leq -b, (b \leq 1)$

$$\log\left(\frac{Z_(t+1)}{Z_t}\right) \leq \log\left(\sum_i \tilde{w}^{(t)}(1-\eta v_{t,1} + \frac{\eta^2}{2}v_{t,i}^2)\right) \tag{1}$$

$$= \log\left(1 - \sum_i \tilde{w}_i^{(t)}\left(\eta v_{t,i} - \frac{\eta^2}{2}v_{t,i}^2\right)\right) \tag{2}$$

$$\leq -\sum_i \tilde{w}_i^{(t)}\left(\eta v_{t,i} - \frac{\eta^2}{2}v_{t,i}^2\right) \tag{3}$$

$$= -\eta < \tilde{w}^{(t)}, v_t > + \frac{\eta^2}{2}\sum_i \tilde{w}_i^{(t)} v_{t,i}^2/2$$

$$\leq -\eta < \tilde{w}^{(t)}, v > + \frac{\eta^2}{2}$$

$$\sum_{t=1}^{T} \log\left(\frac{Z_{t+1}}{Z_t}\right) = \log\left(Z_{T+1}\right) - \log\left(Z_1\right)$$

$$\leq -\eta \sum_{t=1}^{T} < \tilde{w}^{(t)}, v_t > + \frac{T\eta^2}{2}$$

By summing the inequality over t,

$$\therefore \eta \sum_{t=1}^{T} < \tilde{w}^{(t)}, v_t > \quad \leq \quad \log\left(Z_1\right) - \log\left(Z_{t+1}\right) + \frac{T\eta^2}{2}$$

$$\log(Z_1) = \log(d)$$

We lower bound Z_{T+1}. Note that $w_i^{T+1} = e^{-\eta \sum_t v_{t,i}}$,

$$\log\left(Z_{T+1}\right) = \log\left(\sum_i e^{-\eta \sum_t v_{t,i}}\right)$$

$$\geq \log\left(\max_i e^{-\eta \sum_t v_{t,i}}\right)$$

$$= -\eta \min_i \sum_t v_{t,1}$$

With the fact that $\log(Z_1) = \log(d)$,

$$\therefore \eta \sum_{t=1}^{i} < \tilde{w}^{(t)}, v_t > \leq \log(d) + \eta \min_i \sum_t v_{t,i} + \frac{T\eta^2}{2}$$

$$\Rightarrow \sum_{t=1}^{T} < \tilde{w}^{(t)}, v_t > - \min_i \sum_t v_{t,i} \leq \frac{\log(d)}{\eta} + \frac{T\eta}{2}$$

choose $\eta = \sqrt{\frac{2\log(d)}{T}}$, which gives $\leq \sqrt{2\log(d)T}$. □

3 Regret bound for online leaning

Theorem 13. *Let $\mathcal{H} = \{h_1, \cdots, h_d\}$ be a finite hypothesis class. Then Weighted-Majority achieves*

$$\sum_{t=1}^{T} |p_t - y_t| - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(x_t) - y_t| \leq \sqrt{2\log(|\mathcal{H}|)T}$$

Proof. Each experts h_i predicts $h_i(x_t)$ on example x_t. Loss $v_{t,i} = |h_i(x_t) - y_t|$.
The predictions of Weighted-Majority are

$$p_t = \sum_i w_i^{(t)} h_i(x_t).$$

The expected loss is

$$\begin{aligned} |p_t - y_t| &= |\sum_{i=1}^{d} w_i^{(t)} h_i(x_t) - y_t| \\ &= \sum_{i=1}^{d} \left| w_i^{(t)} \left(h_i\left(x_t\right) - y_t \right) \right| \\ &= < w^{(t)}, v_t > \end{aligned}$$

Then we use the regret bound for Weighted-Majority to bound $< w^{(t)}, v_t >$. □

References

[1] Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE, 2020.