Online Learning

Peng Lingwei

August 20, 2019

Contents

21 Online Learning

21.1 ONLINE CLASSIFICATION IN THE REALIZABLE CASE

Online learning is performed in a sequence of consecutive rounds, such as at round t:

- get an instance $\vec{x}_t$;
- predict the instance's label is $\hat{y}_t$;
- get the correct label y_t;
- update the mistakes count.

In the online learning model we make no statistical assumptions regarding the origin of the sequence of examples. It can be deterministic, stochatic, or even adversarially adaptive to the learner's own behavior. If the adversary is arbitrary, the problem is meaningless.So in this section, we require the realizability assumption.

Realizability assumption: the labels are generated by some hypothesis, $h^* : \mathcal{X} \to \mathcal{Y}$. And h^* is in our hyperthesis class $\mathcal{H}$.

Definition 1. ***(Mistake Bounds, Online Learnability).***

$$\forall S_T = ((x_1, h^*(x_1)), \ldots, (x_T, h^*(x_T))), M_A(S_T) = \sum_{i=1}^{T} 1_{[A(S_{i-1})(x_i) \neq h^*(x_i)]}$$

$$M_A(\mathcal{H}) = \sup_{S_T \in (\mathcal{X},\mathcal{Y}) \times \cdots \times (\mathcal{X},\mathcal{Y})} M_A(S_T)$$

In online learning, the goal is to study which hypothesis classes are learnable in the online model.

In ERM rule with realizability assumption, we want $L_{\mathcal{D}}(h_{ERM}(S)) = 0$. Analoguely, we want all rest hypothesises are consistent with all past examples. So, the **Consistent algorithm** maintains a set V_t, of all the hypothesises which are consistent with $((x_1, y_1), \ldots, (x_{t-1}, y_{t-1}))$, which is called the version space.

Corollary 1. *Let $\mathcal{H}$ be a finite hypothesis class, $M_{Consistent}(\mathcal{H}) \leq |\mathcal{H}| - 1$.*

Definition 2. ***(Halving Algorithm).***
Receive x_t, $\hat{y}_t = \arg\max_{y \in \{0,1\}} |\{h \in V_t : h(x_t) = y\}|$. Then update $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$.

Corollary 2. *Let $\mathcal{H}$ be a finite hypothesis class, $M_{Halving}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.*

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

In PAC, any ERM hypothesis is good, but in online learning choosing an arbitrary ERM hypothesis is far from being optimal.

21.1.1 Online Learnability

What's the optimal online learning algorithm for a given hypothesis class $\mathcal{H}$?

A strategy for an adversarial environment can be formally described as a binary tree. A node i is a sample x_i, the left edge means $y_i = h^*(x_i) = 1$, and the right edge means $y_i = y^*(x_i) = 0$.

If the predictor give $\hat{y_i} = 1$, the adversary traverses to the right child, which means $y_i = 0$, otherwise traverses to the left child.

If we code node by line order, then $i_1 = 1$, $i_{t+1} = 2i_t + y_t = 2^{t-1} + \sum_{j=1}^{t-1} y_j 2^{t-1-j}$.

Definition 3. ***($\mathcal{H}$ Shattered Tree).*** *A shattered tree of depth d is a sequence of instance $v_1, \ldots, v_{2^d-1}$ in $\mathcal{X}$ such that for every labeling $(y_1, \ldots, y_d) \in \{0,1\}^d$, there exists $h \in \mathcal{H}$ such that for all $t \in [d]$ we have $h(v_{i_t}) = y_t$ where $i_t = 2^{t-1} + \sum_{j=1}^{t-1} y_j 2^{t-1-j}$.*

Definition 4. ***(Littlestrone's Dimension (Ldim)).*** *$Ldim(\mathcal{H})$ is the maximal integer T such that there exists a shattered tree of depth T, which is shattered by $\mathcal{H}$.*

Lemma 1.

$$\forall A, M_A(\mathcal{H}) \geq Ldim(\mathcal{H}).$$

Definition 5. ***(Standard Optimal Algorithm (SOA)).***

$$\hat{y_t} = \arg \max_{y \in \{0,1\}} Ldim(\{h \in V_t : h(x_t) = y\})$$

Lemma 2.

$$M_{SOA}(\mathcal{H}) \leq Ldim(\mathcal{H}).$$

Proof. We want to get $Ldim(V_{t+1}) \leq Ldim(V_t) - 1$. Because $Ldim(V_{t+1}) \leq Ldim(V_t - V_{t+1})$, then we can construct $Ldim(V_{t+1}) + 1$ deepth tree for class V_t, which is shattered by V_t. Therefore, $Ldim(V_{t+1}) \leq Ldim(V_t) - 1$ □

Theorem 1.

$$\forall \mathcal{H}, VCdim(\mathcal{H}) \leq Ldim(\mathcal{H}).$$

note: $\exists \mathcal{H}, VCdim(\mathcal{H}) = 1, Ldim(\mathcal{H}) = \infty$.

21.2 ONLINE CLASSFICATION IN THE UNREALIZABLE CASE

Definition 6. ***(Regret).***

$$Regret_A(h, T) = \sup_{S_T \in (\mathcal{X},\mathcal{Y}) \times \cdots \times (\mathcal{X},\mathcal{Y})} \left[\sum_{t=1}^{T} |\hat{y_t} - y_t| - \sum_{t=1}^{T} |h(x_t) - y_t| \right]$$

$$Regret_A(\mathcal{H}, T) = \sup_{S_T \in (\mathcal{X},\mathcal{Y}) \times \cdots \times (\mathcal{X},\mathcal{Y})} \left[\sum_{t=1}^{T} |\hat{y_t} - y_t| - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(x_t) - y_t| \right] = \sup_{h \in \mathcal{H}} Regret_A(h, T).$$

We want $Regret_A(\mathcal{H}, T)$ grows sublinearly with the number of rounds, T, which implies that the difference between the error rate of the leaner and the best hypothesis in $\mathcal{H}$ tends to zero as T goes to infinity.

In preceeding section's model, it is impossible. For $\mathcal{H} = \{h_0, h_1\}$, where h_0 always returns 0, and h_1 is always return 1. Then the adversary can make the number of mistakes of any online algorithm be equal to T, but $\inf_{h\in\mathcal{H}} \sum_{t=1}^{T} |h(x_t) - y_t| \leq T/2$, which means that $Regret_A(\mathcal{H}, T) \geq T/2$.

So we need further restrict the power of the adversarial environment. We allow the learner to randomize his predictions, and the adversarial environment only know the probability distribution of the predictions, that is $\mathbb{P}[\hat{y}_t = 1] = p_t$, then

$$\begin{aligned} Regret_A(h, T) &= \sup_{S_T\in(\mathcal{X},\mathcal{Y})\times\cdots\times(\mathcal{X},\mathcal{Y})} \left[\sum_{t=1}^{T} \mathbb{P}\{|\hat{y}_t - y_t|\} - \sum_{t=1}^{T} |h(x_t) - y_t|\right] \\ &= \sup_{S_T\in(\mathcal{X},\mathcal{Y})\times\cdots\times(\mathcal{X},\mathcal{Y})} \left[\sum_{t=1}^{T} \{|p_t - y_t|\} - \sum_{t=1}^{T} |h(x_t) - y_t|\right] \end{aligned}$$

21.2.1 Weighted-Majority

Algorithm 1 Weighted-Majority

Require: $\mathcal{H} = \{h_1, \ldots, h_d\}$; number of rounds T.
Ensure: $\hat{w}^{(1)} = (1, \ldots, 1), \eta = \sqrt{2\log(d)/T}$
  for $t = 1, 2, \ldots$ **do**
    $Z_t = \sum_i^d \hat{w}_i^{(t)}, w^{(t)} = \hat{w}^{(t)}/Z_t$
    Choose hyperthesis h_i at random according to $\mathbb{P}[h_i] = w_i^{(t)}$
    receive costs of all experts $\vec{v}_t \in [0, 1]^d$
    pay cost $\langle \vec{w}^{(t)}, \vec{v}_t \rangle$
    $\forall j, \hat{w}_j^{(t+1)} = \hat{w}_j^t e^{-\eta v_{t,j}}$
  end for.

Here is a abstract concept v_t, which means the cost of each hyperthesis.Let the cost $v_t \in [0, 1]^d$.

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

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

Proof.

$$\log \frac{Z_{t+1}}{Z_t} = \log \sum_{i=1}^{d} \frac{\hat{w}_i^{(t)}}{Z_t} e^{-\eta v_{t,i}} = \log \sum_{i}^{d} w_i^{(t)} e^{-\eta v_{t,i}}.$$

For $e^{-a} \le 1 - a + a^2/2$,

$$\begin{aligned}
\log \frac{Z_{t+1}}{Z_t} &\le \log \sum_{i=1}^{d} w_i^{(t)} \left(1 - \eta v_{t,i} + \eta^2 v_{t,i}^2/2\right) \\
&= \log(1 - \sum_{i=1}^{d} w_i^{(t)} \left(\eta v_{t,i} - \eta^2 v_{t,i}^2/2\right)) \\
&\le - \sum_{i=1}^{d} w_i^{(t)} \left(\eta v_{t,i} - \eta^2 v_{t,i}^2/2\right) \\
&\le - \eta \langle w^{(t)}, v_t \rangle + \eta^2/2 \\
\log(Z_{T+1}) - \log(Z_1) = \sum_{t=1}^{T} \log \frac{Z_{t+1}}{Z_t} &\le -\eta \sum_{t=1}^{T} \langle w^{(t)}, v_t \rangle + \frac{T\eta^2}{2} \\
\log Z_{T+1} = \log \left(\sum_{i=1}^{d} e^{-\eta \sum_t^T v_{t,i}} \right) &\ge \log \left(\max_i e^{-\eta \sum_t^T v_{t,i}} \right) = -\eta \min_i \sum_t^T v_{t,i} \\
-\eta \min_i \sum_t^T v_{t,i} - \log(d) &\le - \eta \sum_{t=1}^{T} \langle w^{(t)}, v_t \rangle + \frac{T\eta^2}{2} \\
\sum_{t=1}^{T} \langle w^{(t)}, v_t \rangle - \min_i \sum_t^T v_{t,i} &\le \frac{\log(d)}{\eta} + \frac{\eta T}{2} \le \sqrt{2 \log(d) T}
\end{aligned}$$

□

Theorem 3.

$\forall h \in \mathcal{H}, \Omega\left(\sqrt{Ldim(\mathcal{H})T}\right) \le Regret_A(h, T) \le \sqrt{2 \min\left\{\log(|\mathcal{H}|), Ldim(\mathcal{H}) \log(eT)\right\} T}$.

Proof. We begin with finite class $\mathcal{H} = \{h_1, \ldots, h_d\}$. The cost $v_{t,i} = |h_i(x_t) - y_t|$.
Let $p_t = \sum_i^d w_i^{(t)} h_i(x_t) \in [0, 1]$, and the loss is

$$|p_t - y_t| = \left| \sum_{i=1}^{d} w_i^{(t)} h_i(x_t) - y_t \right| = \left| \sum_{i=1}^{d} (w_i^{(t)} (h_i(x_i)) - y_t) \right| = \sum_{i=1}^{d} w_i^{(t)} \left| h_i(x_t) - y_t \right| = \langle w^{(t)}, v_t \rangle$$

$$Regret_A(\mathcal{H}, T) \le \sqrt{2 \log(|\mathcal{H}|) T}$$

□

Let $\mathcal{H}$ be any hypothesis class with $Ldim(\mathcal{H}) < \infty$. $\forall \{x_1, x_2, \ldots, x_T\} \in \mathcal{X} \times \cdots \times \mathcal{X}$. For any $h \in \mathcal{H}$, $\exists L \le Ldim(\mathcal{H})$ and indices $1 \le i_1 < i_2 < \cdots < i_L \le T$ such that $Expert(i_1, i_2, \ldots, i_L)$ predicts exactly $h(x_t)$. i_L is the round that SOA algorithm prediction is different with $h(x_t)$. (Seeing expert algorithm.)
The mistake sequences $(i_1, i_2, \ldots, i_L)$'s num is

$$d = \sum_{L=0}^{Ldim(\mathcal{H})} \mathbb{C}_L^T \le \left(\frac{eT}{Ldim(\mathcal{H})} \right)^{Ldim(\mathcal{H})}$$

These experts constructed from $\mathcal{H}$

21.3 ONLINE CONVEX OPTIMIZATION

Convex learning, $\mathcal{H}$ is convex and $\forall z \in Z, l(\cdot, z)$ is a convex function.

$$Regret_A(w,T) = \sum_{t=1}^{T} l(w^{(t)}, z_t) - \sum_{t=1}^{T} l(w, z_t).$$

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

Definition 7. ***(Online Gradient Descent).***

1. $v_t \in \partial_w l(w^{(t)}, z_t)$
2. $w^{(t+1/2)} = w^{(t)} - \eta v_t$
3. $w^{(t+1)} = \arg\min_{w\in\mathcal{H}} \|w - w^{(t+1/2)}\|$

Theorem 4.

$$\forall w \in \mathcal{H}, \quad Regret_A(w,T) \le \frac{\|w\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$

If $\forall t, l(\cdot, z_t)$ is ρ-Lipschitz, then setting $\eta = 1/\sqrt{T}$ yields

$$Regrets_A(w,T) \le \frac{1}{2}(\|w\|^2 + \rho^2)\sqrt{T}$$

If we futher assume that $\mathcal{H}$ is B-bounded and we set $\eta = \frac{B}{\rho\sqrt{T}}$, then

$$Regret_A(\mathcal{H},T) \le B\rho\sqrt{T}$$

Proof.

$$\begin{aligned}
&\|w^{(t+1)} - w\|^2 - \|w^{(t)} - w\|^2 = \|w^{(t+1/2)} - w\|^2 - \|w^{(t)} - w\|^2 \\
&= -2\eta\langle w^{(t)} - w, v_t\rangle + \eta^2\|v_t\|^2 \le -2\eta(l(w^{(t)}, z_t) - l(w, z_t)) + \eta^2\|v_t\|^2 \\
&\|w^{(T+1)} - w\|^2 - \|w^{(1)} - w\|^2 \le -2\eta\sum_{t=1}^{T}(l(w^{(t)}, z_t) - l(w, z_t)) + \eta^2\sum_{t=1}^{T}\|v_t\|^2 \\
&Regret_A(w,T) \le \frac{\|w\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2
\end{aligned}$$

□

In previous proof, η is dependented on T. We can use doubling trick to avoid this. We let $2^M \le T < 2^{M+1}$. First, we have $Regret_A(w, 2^m) \le \alpha\sqrt{2^m}$.

$$Regret_A(w,T) \le \sum_{m=1}^{M+1} Regret_A(w, 2^m) \le \alpha\sum_{m=1}^{M+1}(\sqrt{2})^m \le \alpha\frac{(\sqrt{2})^{M+2} - \sqrt{2}}{\sqrt{2}-1} = \frac{2}{\sqrt{2}-1}\alpha\sqrt{T}.$$

21.4 THE ONLINE PERCEPTRON ALGORITHM

The perceptron's hypothesis class of homogenous halfspaces $\mathcal{H} = \{x \mapsto sign(\langle w, x\rangle)\}$, If $d \geq 2$ then $Ldim(\mathcal{H}) = \infty$.

The perceptron equals to gradient descent for calculating minimization of $l^{hinge} = \max\{0, 1 - y_t\langle w, x_t\rangle\}$.The subgradient is

$$v_t = -1_{[y_t\langle w^{(t)}, x_t\rangle \leq 0]} y_t x_t$$

Using preceeding section theorem, we have

$$\sum_{t=1}^{T} l^{hinge}(w^{(t)}, z_t) - \sum_{t=1}^{T} l^{hinge}(w, z_t) \leq \frac{1}{2\eta}\|w\|_2^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|_2^2$$

Let $M = \sum_{t=1}^{T} l^{0-1}(w^{(t)}, z_t)$, then

$$M - \sum_{t=1}^{T} l^{hinge}(w, z_t) \leq \frac{1}{2\eta}\|w\|_2^2 + \frac{\eta}{2}MR^2 = R\|w\|\sqrt{M}, \quad \eta = \frac{\|w\|}{R\sqrt{M}}, R = \max_t \|x_t\|$$

$$M \leq \sum_{t=1}^{T} l^{hinge}(w, z_t) + R\|w\|\sqrt{\sum_{t=1}^{T} l^{hinge}(w, z_t) + R^2\|w\|^2}$$

If $\exists w^*, \forall t, y_t\langle w^*, x_t\rangle \geq 1$, then $M \leq R^2\|w^*\|^2$. (Separability with large margin)

$(x - b\sqrt{x} - c \leq 0 \Rightarrow x \leq c + b^2 + b\sqrt{c})$

note: there can be cases in which there exists some w^* that makes zero errors on the sequence but the Perceptron will make many errors, which is a direct consequence of the fact that $Ldim(\mathcal{H}) = \infty$. We sidestep this impossibility result is assuming that adding assumption that $\sum_{t=1}^{T} l^{hinge}(w, z_t)$ is note excessively large.