IE613: Online Machine Learning **Jan-Apr 2016**

Lecture 7: Lower Bound of Online Learning Algorithm

Lecturer: M. K. Hanawal *Scribes: Nithilaksh P L*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In the previous class we saw that for any $\mathcal{H}$ such that $L_{dim}(\mathcal{H}) < \infty$ and $1 \leq i_1 < \cdots < i_L \leq T$ and $L_{dim}(\mathcal{H})$, an algorithm can be used to convert the hypothesis class into a finite one. Then a Weighted Majority algorithm may be used. Furthermore, we proved that for all sequences,

$$\mathbb{E}[R_{\mathcal{H}}](T, WM)] \leq \sqrt{2L_{dim}(\mathcal{H})T}$$

Here we will now show that there exists a sequence $(x_1, y_1)\cdots(x_T, y_T)$ such that,

$$\sum_{t=1}^{T} |\hat{y}_t - y_t| - \min_{h\in\mathcal{H}} |h(x_t) - y_t| \geq \sqrt{\frac{L_{dim}(\mathcal{H})T}{8}}$$

Lower bound: Let $\mathcal{H}$ be any hypothesis class such that $L_{dim}(\mathcal{H}) < \infty$ for any algorithm there exists a sequence $(x_1, y_1)\cdots(x_T, y_T)$ such that $\sum_{t=1}^{T} |\hat{y}_t - y_t| - \min_{h\in\mathcal{H}} |h(x_t) - y_t| \geq \sqrt{\frac{L_{dim}(\mathcal{H})T}{8}}$.

Lemma 7.1 *(Kichine's Inequality) If $\{\sigma_i\}_{i=1}$ are a sequence of i.i.d random variables such that $P(\{\sigma_i = 1\}) = P(\{\sigma_i = -1\})\frac{1}{2}$ then, $\mathbb{E}[|\sum_{i=1}^{T} \sigma_i a_i|] \geq \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{T} a_i^2}$ where a_i are real numbers.*

Proof: Let $d = L_{dim}(\mathcal{H})$, $k = T/d$; T is such that k is an integer. Consider a shattered tree $(v_1, v_2, \cdots v_{2^{d-1}}$ for $\mathcal{H}$. We will construct a sequence $(x_1, y_1)\cdots(x_T, y_T)$ by following a path $(u_1, z_1), \cdots(u_d, z_d)$.

Based on $z_i \in \{0, 1\}$ determine to move left or right in the tree. Each node u_i, $i = 1, \cdots, d$ dcorresponds to a block $(x_{(i-1)k+1}, y_{(i-1)k+1}), x_{(i-1)k+2}, y_{(i-1)k+2}), \cdots, (x_{ik}, y_{ik})$ where $x_{(i-1)k+1} = x_{(i-1)k+2} = \cdots = x_{ik} = u_i$ and $y_{(i-1)k+1}, y_{(i-1)k+2}, \cdots, y_{ik}$ are chosen independently and uniformly over $\{0, 1\}$.

Let $T_i = \{(i-1)k+1, (i-1)k+2, \cdots, ik\}$, $T = T_1 T_2 \cdots T_d$. Let $r_i = \sum_{t\in T_i} y_t$ and $Z_i = \arg\min_{\bar{Z}_i\in\{0,1\}} \sum_{t\in T_i} |\bar{z}_t - y_t|$. Now,

$$\begin{aligned}
\min_{\bar{Z}_i\in\{0,1\}} \sum_{t\in T_i} |\bar{z}_t - y_t| &= \begin{cases} r_i & \text{if } r_i < k/2 \\ k - r_i & \text{else} \end{cases} \\
&= \begin{cases} k/2 - r_i & \text{if } r_i < k/2 \\ -k/2 + r_i & \text{else} \end{cases} \\
&= \left| r_i - \frac{k}{2} \right|
\end{aligned}$$

Take z_i using $\arg\min_{\bar{Z}_i\in\{0,1\}} \sum_{t\in T_i} |\bar{z}_t - y_t|$. Now there exists $h \in \mathcal{H}$ such that $h(u_i) = z_i$.

$$\frac{k}{2} - \min_{\bar{Z}_i\in\{0,1\}} \sum_{t\in T_i} |\bar{z}_t - y_t| = \left| r_i - \frac{k}{2} \right|$$

Summing over all d blocks and taking the expectation,

$$\frac{dk}{2} - \mathbb{E} \min_{\bar{Z}_i \in \{0,1\}} \sum_{t \in T_i} |\bar{z}_t - y_t| = \sum_{i=1}^{d} \mathbb{E}\left| r_i - \frac{k}{2} \right|$$

$$\mathbb{E}\left[\left| r_i - \frac{k}{2} \right|\right] = \mathbb{E}\left[\left| \sum_{t \in T_i} y_t - \frac{k}{2} \right|\right]$$

$$= \frac{1}{2}\mathbb{E}\left[\left| \sum_{t \in T_i} (2y_t - 1) \right|\right]$$

$$\geq \frac{1}{2} \times \frac{1}{\sqrt{2}} \sqrt{|T_I|} = \sqrt{\frac{T}{8d}}$$

To complete the proof we have to argue that $\frac{dk}{2} \leq \mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right]$. Since y_t is uniformly chosen,

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] = \frac{T}{2} = \frac{kd}{2}$$

■

7.1 A Variant with Noisy Labels

Consider the following version of the realizable case. The labels are generated according to a fixed hypothesis in each round, but we get to observe only noisy version. Specifically, fixed hypothesis $h \in \mathcal{H}$ and $y_t = h(x_t) \oplus \nu_t$ where ν_t is the Bernoulli random variable with parameter $< \frac{1}{2}$ we see a flipped version but we don't know when the flipping is done.

Theorem 7.2 $\mathcal{H}, \gamma \in [0, \frac{1}{2}]$ $\exists$ *an Online Learning Algorithm such that for any* $h \in \mathcal{H}$ *and any labels* $(x_1, y_1), \cdots, (x_T, y_T)$ *where* $P(\{y_t \neq h(x_t)\}|x_t) = \nu$ *then,*

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - h(x_t)|\right] - \mathbb{E}\left[\min_{g \in \mathcal{H}} \sum_{t=1}^{T} |g(x_t) - h(x_t)|\right] \leq \frac{L_{dim}(\mathcal{H}) \log(T)}{2\sqrt{\gamma(1-\gamma)}}$$

Thus, this is a stochastic variant of the online learning setting. As seen, the regret bound now scales as $\log(T)$, as exponential improvement compared to $\sqrt{T}$ in the non-stochastic (adversarial) case.