IE613: Online Machine Learning **Jan-Apr 2018**

Lecture 5: Littlestone's Dimension

Lecturer: M. K. Hanawal *Scribes: Harsh Sheth*

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.*

5.1 Recap

Online-Learnability
In the last class, we proved that the weighted majority algorithm gives a sub-linear upper bound for expected regret for finite hypothesis classes in the unrealizable (agnostic) case.

$$\mathbb{E}[R_{\mathcal{H}}(WM, T)] \leq \sqrt{2log(|\mathcal{H}|).T} \tag{5.1}$$

Therefore, finite hypothesis classes are online-learnable in the agnostic case.

In this class, we continue our discussion on **Online Learnability** as we define the Littlestone's dimension or the Little dimension (Ldim($\mathcal{H}$)) and see some examples.

5.2 Littlestone's Dimension

Similar to the VC-Dimension for the supervised learning case, Littlestones's dimension tries to measure the complexity or learnability of a hypothesis class. It will be shown that the upper bound for expected regret is actually a function of the Little dimension and not the cardinality of a hypothesis class. Therefore, a hypothesis class with finite Little dimension, even if it has an infinite size, will be shown to be learnable.

5.2.1 Strategy for Adversarial Environment

To motivate Little dimension, we view the online learning process as a game played between the learner and the environment. The objective of the environment, is to make the learner err for the first T rounds, for as large a T as possible. However, there is a constraint on the environment, as we re-introduce the realizability assumption, i.e. the environment's strategy in every time instance must exactly correspond with at least one hypothesis $h^* \in \mathcal{H}$.

The environment does not select an h^* a priori, instead the strategy of the environment is to declare the learner's prediction to be wrong for as many rounds as possible as long as at least one hypothesis from the hypothesis class agrees with all the outputs provided by the environment, i.e., $\exists\ h^* \in \mathcal{H}$ s.t. $y_t = h^*(x_t)\ for\ t = 1, 2, ..., T$. Remember, the environment also controls what instances are provided to the learner, and the order that they are provided in.

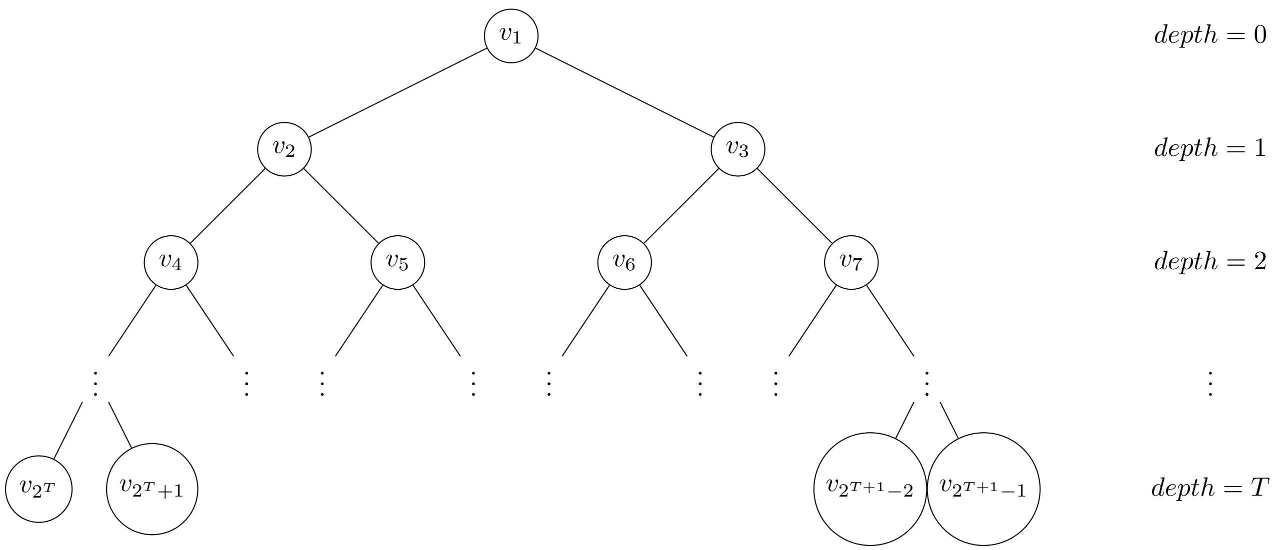

Figure 5.1: General strategy tree for the environment

The environment's optimal strategy in this situation can formally described as a binary tree. Every node in the tree is associated with an instance. At every round, the environment declares the learner's prediction to be wrong and provides the next instance to the learner in the following manner:

1. Initially present the instance associated with the root of the tree to the learner
2. If the learner predicted $\hat{y}_t = 0$,
   (a) declare $y_t = 1$
   (b) go to right child
3. If the learner predicted $\hat{y}_t = 1$,
   (a) declare $y_t = 0$
   (b) go to left child
4. Present instance associated with new node to the learner

The binary tree takes the form show in Fig 5.1. It is a complete binary tree of depth T. Such a tree will have $2^{T+1} - 1$ nodes. Assume there is an instance (data point from $\mathcal{X}$) attached to each node, and let the instances be called $v_1, v_2, \ldots, v_{2^{T+1}-1}$. Let i_t be the index of the node that the environment reaches at round t. Then,

$$\begin{aligned} i_{t+1} &= 2i_t + y_t \\ &= 2^t + \sum_{j=1}^{t} y_j 2^{t-j} \end{aligned}$$

Definition 5.1 ($\mathcal{H}$ shattered tree) *A shattered tree of depth d is a sequence of instances $v_1, v_2, \ldots, v_{2^d-1}$ in X such that for all labellings $(y_1, y_2, \ldots, y_d) \in \{0,1\}^d$, $\exists\ h \in \mathcal{H}$ such that $\forall t \in [d]$ we have $h(v_t) = y_t$.*

Note: For a shattered tree of depth d, only the nodes of the first $d-1$ nodes are relevant, this is because $(y_1, y_2, \ldots, y_d)$ are independent of the leaf nodes.

In Fig 5.3, labeling $\{y_1, y_2\} = \{0, 0\}$ is achieved by hypothesis h_1, $\{0, 1\}$ by h_2, $\{1, 0\}$ by h_3 and $\{1, 1\}$ by h_4. Therefore, it is shattered by the given hypothesis class. The optimal strategy for the environment described previously, only works on a on an $\mathcal{H}$ shattered tree.

	h_1	h_2	h_3	h_4
v_1	0	0	1	1
v_2	0	1	$*$	$*$
v_3	$*$	$*$	0	1

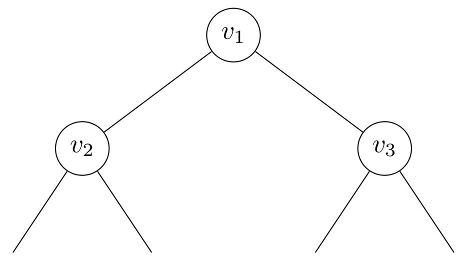

Figure 5.2: An example of a shattered tree of depth $d = 2$

Definition 5.2 (Littlestone's dimension (LDim)) *LDim($\mathcal{H}$) is the maximal integer d such that there exists a shattered tree of depth d which is shattered by $\mathcal{H}$.*

The Little dimension defines the maximum depth d upto which the environment can use its optimal strategy.

Example 1:
Let $\mathcal{X} = \{x_1, x_2, \ldots, x_d\}$, $\mathcal{H}^{(1)} = \{h_1, h_2, \ldots, h_d\}$. $h_j(x) = 1$ iff $x = x_j$.

	h_1	h_2	$\cdots$	h_d
$v1$	1	0	$\cdots$	0
v_2	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
v_d	0	0	$\cdots$	1

We try to calculate the Little dimension for this hypothesis class $\mathcal{H}^{(1)}$.

For $d = 1$, shattered tree requires a single node. If label is 1, any hypothesis h_i shatters the tree for instance v_i, while for label 0, all hypotheses except h_i shatter the tree for instance v_i. For $d = 2$, 3 instances for three nodes are required. As no hypothesis predicts the label 1 at more than 1 point, the labeling $\{1, 1\}$ can be achieved by no hypothesis. Therefore, LDim($\mathcal{H}^{(1)}$) = 1.

Example 2:
Let $X = [0, 1]$ and $\mathcal{H}^{(2)} = \{x \mapsto \mathbb{1}_{[x<a]} : a \in [0, 1]\}$. For this hypothesis class, a shattered tree of depth d, for any positive real number d is given in Fig 5.3. Thus, LDim($\mathcal{H}^{(2)}$) = ∞.

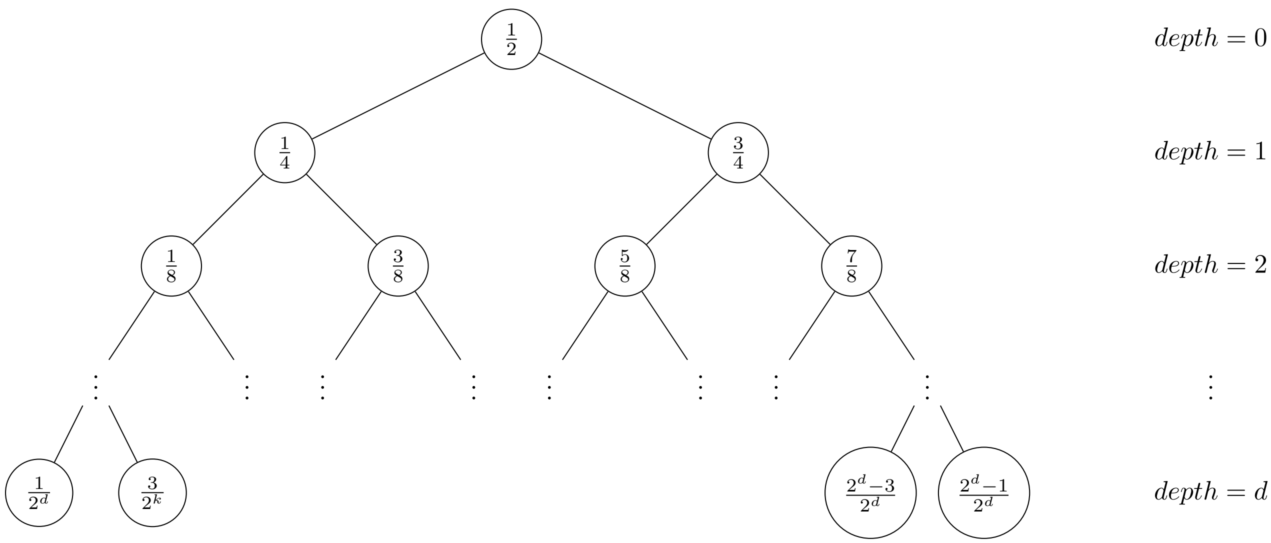

Figure 5.3: Shattered tree for $\mathcal{H}^{(2)}$

Illustration for $d = 3$:

- $a \in [0, 1]$
- $y_1 = 0 \Longrightarrow a \in \left[0, \frac{1}{2}\right]$
- $y_2 = 1 \Longrightarrow a \in \left(\frac{1}{4}, \frac{1}{2}\right]$
- $y_3 = 0 \Longrightarrow a \in \left(\frac{1}{4}, \frac{3}{8}\right]$

Therefore, there is always a range of values of a for which the outcomes declared by the environment correspond with a hypothesis in $\mathcal{H}^{(2)}$