

Littlestone Dimension

Optimal Mistake Bounds in Online Prediction

Learning Theory

Online Prediction

In deterministic, realizable online prediction, the optimal mistake bound is exactly the Littlestone dimension. This note defines and proves this characterization, then compares it with VC dimension: real thresholds, intervals, and affine halfspaces all have finite VC dimension but no finite mistake bound.

1. Online prediction and realizability

Let $\mathcal{X} \neq \emptyset$ be an input space and $\emptyset \neq \mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ a class of binary-valued functions. Write $\mathbb{N}_0 := \{0, 1, 2, \dots\}$.

A deterministic online learner is a sequence of maps $A = (A_t)_{t \geq 1}$, where

$$A_t : (\mathcal{X} \times \{0, 1\})^{t-1} \times \mathcal{X} \longrightarrow \{0, 1\}.$$

At round t , the learner observes x_t , predicts $\hat{y}_t = A_t(z_{<t}, x_t)$, and then receives the label y_t ; here $z_{<t} := ((x_s, y_s))_{s < t}$ records the preceding inputs and labels.

Definition 1 Realizability

A sequence $z_{1:T} := ((x_t, y_t))_{t=1}^T$ is $\mathcal{H}$ -realizable if there exists $h^* \in \mathcal{H}$ such that

$$y_t = h^*(x_t), \quad t = 1, \dots, T.$$

Let $\mathcal{S}_T(\mathcal{H})$ denote the set of these sequences.

Define the number of mistakes and its worst-case bound by

$$M_T(A; z_{1:T}) := \sum_{t=1}^T \mathbf{1}\{\hat{y}_t \neq y_t\},$$

$$M_A(\mathcal{H}) := \sup_{T \geq 1} \sup_{z_{1:T} \in \mathcal{S}_T(\mathcal{H})} M_T(A; z_{1:T}).$$

The optimal mistake bound is

$$M^*(\mathcal{H}) := \inf_A M_A(\mathcal{H}),$$

where the infimum ranges over all deterministic online learners, with no computational restrictions; both t and T range over the positive integers.

The question is when $M^*(\mathcal{H}) < \infty$: can a constant independent of the horizon T bound the number of mistakes on every $\mathcal{H}$ -realizable sequence? The trivial bound $M_T \leq T$ does not meet this requirement.

2. Trees and shattering

A complete binary tree of depth $d \in \mathbb{N}_0$, with internal nodes taking values in $\mathcal{X}$, can be represented as

$$\mathbf{x} = (\mathbf{x}_t)_{t=1}^d, \quad \mathbf{x}_t : \{0, 1\}^{t-1} \longrightarrow \mathcal{X}.$$

Each internal node specifies an input, and each edge a label. The node reached by following the labels $y_{<t} := (y_1, \dots, y_{t-1})$ carries the input $\mathbf{x}_t(y_{<t})$; $y_{<1}$ is the empty sequence.

Definition 2 Littlestone dimension

The class $\mathcal{H}$ shatters $\mathbf{x}$ if, for every $y \in \{0, 1\}^d$, the path

$$((\mathbf{x}_t(y_{<t}), y_t))_{t=1}^d$$

is $\mathcal{H}$ -realizable (Definition 1). Define

$$\text{Ldim}(\mathcal{H}) := \sup \{d \in \mathbb{N}_0 : \mathcal{H} \text{ shatters some tree of depth } d\}.$$

Every nonempty class shatters the depth-zero tree; we also set $\text{Ldim}(\emptyset) := -1$.

3. The optimal mistake bound

The following theorem is due to Littlestone [1]; see [2, Theorem 10] for the formulation for arbitrary function classes.

Theorem 1 Littlestone, 1988

For every nonempty class $\mathcal{H}$ of binary-valued functions,

$$M^*(\mathcal{H}) = \text{Ldim}(\mathcal{H}).$$

Lower bound

Fix a deterministic learner A and a tree of depth $d \geq 1$ shattered by $\mathcal{H}$. For $t = 1, \dots, d$, set

$$x_t := \mathbf{x}_t(y_{<t}), \quad \hat{y}_t := A_t(z_{<t}, x_t), \quad y_t := 1 - \hat{y}_t.$$

By Definition 2,

$$\exists h \in \mathcal{H}, \quad \forall t \leq d, \quad h(x_t) = h(\mathbf{x}_t(y_{<t})) = y_t.$$

Thus $z_{1:d} \in \mathcal{S}_d(\mathcal{H})$ and $M_A(\mathcal{H}) \geq M_d(A; z_{1:d}) = d$. Taking the infimum over A and then the supremum over the depths of trees shattered by $\mathcal{H}$ gives $M^*(\mathcal{H}) \geq \text{Ldim}(\mathcal{H})$. When $\text{Ldim}(\mathcal{H}) = 0$, the bound follows from nonnegativity.

Upper bound: SOA

Fix any $\mathcal{H}$ -realizable sequence and suppose $D := \text{Ldim}(\mathcal{H}) < \infty$. At round t , retain the functions consistent with the observed labels to form the version space V_t . Partition it into two subclasses according to their values at x_t :

$$V_t := \{h \in \mathcal{H} : h(x_s) = y_s \text{ for all } s < t\},$$

$$V_t^{(b)} := \{h \in V_t : h(x_t) = b\}, \quad b \in \{0, 1\}.$$

After observing y_t , set $V_{t+1} = V_t^{(y_t)}$.

Let $k := \text{Ldim}(V_t)$. Both subclasses have dimension at most k , and

$$\min_{b \in \{0,1\}} \text{Ldim}(V_t^{(b)}) \leq k - 1.$$

Otherwise, each subclass would shatter a tree of depth k . Attaching these two trees below a root x_t would yield a tree of depth $k + 1$ shattered by V_t , contradicting the definition of k .

The Standard Optimal Algorithm (SOA) predicts

$$\hat{y}_t \in \arg \max_{b \in \{0,1\}} \text{Ldim}(V_t^{(b)}),$$

with ties always broken in favor of 0.

Remark Computability

The dimensions of the two subclasses are well-defined, so fixing the tie-breaking rule makes SOA a prediction map. Defining this map does not compute it: no method for computing these two dimensions is given here, nor is a computable implementation of SOA guaranteed [2].

If the prediction is wrong, the retained branch is the other one, whose dimension is the smaller of the two:

$$\text{Ldim}(V_{t+1}) = \min_{b \in \{0,1\}} \text{Ldim}(V_t^{(b)}) \leq \text{Ldim}(V_t) - 1.$$

If the prediction is correct, $V_{t+1} \subseteq V_t$ ensures that the dimension does not increase. Each mistake lowers the dimension by at least one, while realizability ensures $V_{T+1} \neq \emptyset$, so its dimension cannot fall below zero. Since $V_1 = \mathcal{H}$, the total number of mistakes satisfies

$$M_T(\text{SOA}; z_{1:T}) \leq D - \text{Ldim}(V_{T+1}) \leq D.$$

Thus $M^*(\mathcal{H}) \leq D$, which, together with the lower bound, proves the result. When $\text{Ldim}(\mathcal{H}) = \infty$, the lower bound alone proves the result. $\square$

4. Comparison with VC dimension

Definition 3 VC dimension

The class $\mathcal{H}$ shatters a finite set $F \subseteq \mathcal{X}$ if

$$\forall \ell : F \rightarrow \{0, 1\}, \quad \exists h \in \mathcal{H}, \quad \forall x \in F, \quad h(x) = \ell(x).$$

The VC dimension (Vapnik–Chervonenkis dimension) is

$$\text{VCdim}(\mathcal{H}) := \sup\{|F| : F \subseteq \mathcal{X} \text{ finite, } \mathcal{H} \text{ shatters } F\}.$$

Every nonempty class shatters the empty set. Both $\text{VCdim}(\mathcal{H})$ and $\text{Ldim}(\mathcal{H})$ take values in $\mathbb{N}_0 \cup \{\infty\}$.

Proposition 1 Comparison of dimensions

$$\text{VCdim}(\mathcal{H}) \leq \text{Ldim}(\mathcal{H}), \quad \text{Ldim}(\mathcal{H}) \leq \lfloor \log_2 |\mathcal{H}| \rfloor \quad (\mathcal{H} \text{ finite}).$$

Proof. If $\mathcal{H}$ shatters $\{u_1, \dots, u_d\}$, it also shatters the tree $\mathbf{x}_t(y_{<t}) = u_t$, giving the first inequality. A function realizes at most one path in a fixed tree. Shattering a tree of depth d therefore requires at least 2^d functions, giving the second inequality. $\square$

In VC shattering, the inputs are fixed in advance; in a tree, they may depend on the preceding labels.

Real threshold functions

Let $h_\theta(x) := \mathbf{1}\{x \leq \theta\}$, with $x, \theta \in \mathbb{R}$. This class shatters a single point but cannot assign labels $(0, 1)$ to two increasing inputs, so its VC dimension is 1. To determine its Ldim, it suffices to restrict the parameter to $(0, 1)$. Given a feasible parameter interval (a, b) , query its midpoint $x = (a + b)/2$ and retain

$$y = 0 \implies \text{retain } (a, x), \quad y = 1 \implies \text{retain } (x, b).$$

Either label leaves a nonempty open interval. For every finite depth, this constructs a tree shattered by the class. Hence [3, §21.1.1]

$$\text{VCdim}(\{h_\theta : \theta \in \mathbb{R}\}) = 1, \quad \text{Ldim}(\{h_\theta : \theta \in \mathbb{R}\}) = \infty.$$

5. Two familiar examples

Intervals and affine halfspaces also have finite VC dimension and infinite Ldim. Both reproduce the threshold construction above.

Intervals. On the input space $\mathbb{R}$, define

$$\mathcal{H}_{\text{int}} := \{x \mapsto \mathbf{1}\{a \leq x \leq b\} : a, b \in \mathbb{R}, a \leq b\}.$$

This class shatters any two distinct points but cannot realize the labels $(1, 0, 1)$ on three increasing points. This determines its VC dimension; the Ldim claim is proved below:

$$\text{VCdim}(\mathcal{H}_{\text{int}}) = 2, \quad \text{Ldim}(\mathcal{H}_{\text{int}}) = \infty.$$

Affine halfspaces. For an integer $p \geq 1$, take the input space $\mathbb{R}^p$ and define

$$\mathcal{H}_{\text{hs}}^{(p)} := \{x \mapsto \mathbf{1}\{w^\top x + b \geq 0\} : w \in \mathbb{R}^p, b \in \mathbb{R}\}.$$

The classical VC dimension result [3, §9.1.3] and the construction below give

$$\text{VCdim}(\mathcal{H}_{\text{hs}}^{(p)}) = p + 1, \quad \text{Ldim}(\mathcal{H}_{\text{hs}}^{(p)}) = \infty.$$

No margin condition is imposed here; a finite parameter dimension does not guarantee a finite mistake bound.

Proof of the Ldim claims. Let $t, \theta \in (0, 1)$.

For the interval $[0, \theta]$, the label of input t is $\mathbf{1}\{t \leq \theta\}$.

For the halfspace $x_1 \leq \theta$, query only points $(t, 0, \dots, 0)$ on the first coordinate axis. The label is again $\mathbf{1}\{t \leq \theta\}$.

In both cases, the threshold bisection tree from the preceding section applies directly, so both classes have infinite Ldim.

References

- [1] N. Littlestone (1988). [Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm.](#) *Machine Learning*, 2:285–318.
- [2] N. Hasrati and S. Ben-David (2023). [On Computable Online Learning.](#) *ALT*, PMLR 201:707–725.
- [3] S. Shalev-Shwartz and S. Ben-David (2014). [Understanding Machine Learning: From Theory to Algorithms.](#) Cambridge University Press.