

Littlestone Dimension

在线预测中的最优错误界

学习理论

在线预测

在确定性、realizable 在线预测中，最优错误界恰好等于 Littlestone dimension。本文给出这一刻画的定义与证明，并分析它与 VC dimension 的区别：实数阈值、区间和仿射半空间的 VC dimension 都有限，却没有有限的 mistake bound。

1. 在线预测与 realizability

设 $\mathcal{X} \neq \emptyset$ 为输入空间， $\emptyset \neq \mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ 为一类二值函数。记 $\mathbb{N}_0 := \{0, 1, 2, \dots\}$ 。

一个确定性在线学习器是映射序列 $A = (A_t)_{t \geq 1}$ ，其中

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

第 t 轮，学习器观察 x_t ，预测 $\hat{y}_t = A_t(z_{<t}, x_t)$ ，再收到标签 y_t ；其中 $z_{<t} := ((x_s, y_s))_{s < t}$ 记录此前的输入和标签。

定义 1 Realizability

称序列 $z_{1:T} := ((x_t, y_t))_{t=1}^T$ 为 $\mathcal{H}$ -realizable，若存在 $h^* \in \mathcal{H}$ ，使得

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

记这类序列的集合为 $\mathcal{S}_T(\mathcal{H})$ 。

定义错误次数及其最坏情形界：

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

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

最优错误界定义为

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

其中下确界遍历所有确定性在线学习器，不施加计算限制； t, T 均取正整数。

问题是何时 $M^*(\mathcal{H}) < \infty$ 。也就是能否用一个与轮数 T 无关的常数，控制所有 $\mathcal{H}$ -realizable 序列上的错误次数。平凡界 $M_T \leq T$ 不满足这一要求。

2. 树与 shattering

深度为 $d \in \mathbb{N}_0$ 、内部节点取值于 $\mathcal{X}$ 的完全二叉树可表示为

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

每个内部节点对应一个输入，每条边对应一个标签。沿标签 $y_{<t} := (y_1, \dots, y_{t-1})$ 到达的节点，其输入为 $\mathbf{x}_t(y_{<t})$ ； $y_{<1}$ 为空序列。

定义 2 Littlestone dimension

若对每个 $y \in \{0, 1\}^d$ ，路径

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

都是 $\mathcal{H}$ -realizable (定义 1)，则称 $\mathcal{H}$ shatters $\mathbf{x}$ 。定义

$$\text{Ldim}(\mathcal{H}) := \sup\{d \in \mathbb{N}_0 : \mathcal{H} \text{ 能 shatter 某棵深度为 } d \text{ 的树}\}.$$

非空类总能 shatter 深度零的树；另约定 $\text{Ldim}(\emptyset) := -1$ 。

3. 最优错误界

下面的定理来自 Littlestone [1]；任意函数类的表述见 [2, Theorem 10]。

定理 1 Littlestone, 1988

对任意非空二值函数类 $\mathcal{H}$ ，

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

下界

固定确定性学习器 A ，取一棵 $\mathcal{H}$ 能 shatter 的深度 $d \geq 1$ 的树。对 $t = 1, \dots, d$ ，令

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

由定义 2，

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

故 $z_{1:d} \in \mathcal{S}_d(\mathcal{H})$ ，且 $M_A(\mathcal{H}) \geq M_d(A; z_{1:d}) = d$ 。先对 A 取下确界，再对 $\mathcal{H}$ 能 shatter 的树深度取上确界，得到 $M^*(\mathcal{H}) \geq \text{Ldim}(\mathcal{H})$ ； $\text{Ldim}(\mathcal{H}) = 0$ 时由非负性成立。

上界：SOA

固定任意 $\mathcal{H}$ -realizable 序列，并设 $D := \text{Ldim}(\mathcal{H}) < \infty$ 。第 t 轮，保留与已知标签一致的函数，得到 version space V_t 。再按它们在 x_t 上的取值分成两个子类：

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

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

观察 y_t 后，令 $V_{t+1} = V_t^{(y_t)}$ 。

令 $k := \text{Ldim}(V_t)$ 。两个子类的维数都不超过 k ，且

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

否则，两个子类各有一棵深度 k 的树可被 shatter。将它们接到根 x_t 下，就得到一棵 V_t 能 shatter 的深度 $k + 1$ 的树，与 k 的定义矛盾。

Standard Optimal Algorithm (SOA) 预测

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

平局时固定预测 0。

注 可计算性

两个子类的维数是确定的，固定平局规则后，SOA 就定义了一个预测映射。但定义不等于计算：这里未给出计算这两个维数的方法，也不保证 SOA 有可计算的实现 [2]。

若预测错误，留下的是另一个分支，其维数等于两者的最小值：

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

若预测正确，由 $V_{t+1} \subseteq V_t$ ，维数不会增加。每次错误使维数至少下降一，而 realizability 保证 $V_{T+1} \neq \emptyset$ ，维数不会降到零以下。再由 $V_1 = \mathcal{H}$ ，累计错误次数满足

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

故 $M^*(\mathcal{H}) \leq D$ ，结合下界即得结论。 $\text{Ldim}(\mathcal{H}) = \infty$ 时，下界已给出结论。 $\square$

4. 与 VC dimension 的比较

定义 3 VC dimension

对有限集 $F \subseteq \mathcal{X}$ ，若

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

则称 $\mathcal{H}$ shatters F 。VC dimension (Vapnik–Chervonenkis dimension) 定义为

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

非空类总能 shatter 空集。 $\text{VCdim}(\mathcal{H})$ 与 $\text{Ldim}(\mathcal{H})$ 均取值于 $\mathbb{N}_0 \cup \{\infty\}$ 。

命题 1 维数的比较

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

证明. 若 $\mathcal{H}$ shatters $\{u_1, \dots, u_d\}$ ，则它也 shatters 树 $\mathbf{x}_t(y_{<t}) = u_t$ ，得到第一个不等式。

一个函数在固定树中至多实现一条路径。故要 shatter 深度 d 的树，至少需要 2^d 个函数，得到第二个不等式。 $\square$

VC shattering 的输入预先固定；树中的输入则可以依赖此前的标签。

实数阈值函数

令 $h_\theta(x) := \mathbf{1}\{x \leq \theta\}$, $x, \theta \in \mathbb{R}$ 。这个类能 shatter 单点，却不能给两个递增的输入分别标记 $(0, 1)$ ，故 VC dimension 为 1。要看它的 Ldim，只需将参数限制在 $(0, 1)$ 。给定参数的可行区间 (a, b) ，查询中点 $x = (a + b)/2$ ，则

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

无论标签为何，都能保留一个非空开区间。对每个有限深度，都能这样构造一棵被该类 shatter 的树。因此 [3, §21.1.1]

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

5. 两个常见例子

区间与仿射半空间也有有限 VC dimension、无限 Ldim。两者都能重现前节的阈值构造。区间。输入空间为 $\mathbb{R}$ ，定义

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

这个类能 shatter 任意两个不同的点，但不能在三个递增的点上实现标签 $(1, 0, 1)$ 。这给出 VC dimension；Ldim 的结论在下文证明：

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

仿射半空间。给定整数 $p \geq 1$ ，输入空间为 $\mathbb{R}^p$ ，定义

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

VC dimension 的经典结果 [3, §9.1.3] 与下文的构造给出

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

这里不要求分类间隔；参数维数有限，并不保证 mistake bound 有限。

Ldim 的证明。令 $t, \theta \in (0, 1)$ 。

取区间 $[0, \theta]$ ，则输入 t 的标签为 $\mathbf{1}\{t \leq \theta\}$ 。

取半空间 $x_1 \leq \theta$ ，只查询第一坐标轴上的点 $(t, 0, \dots, 0)$ 。标签同样是 $\mathbf{1}\{t \leq \theta\}$ 。

两种情况下，前节的阈值二分树都可直接使用，因而 Ldim 都是无穷。

参考文献

- [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.