輪講資料 Understanding Machine Learning: From Theory to Algorithms Part IV

Masanari Kimura

June 14, 2019

Abstract

本資料は書籍”Understanding Machine Learning: From Theory to Algorithms”[1] の輪講資料です．本資料は該当書籍の Chapter5 の内容を含みます．

1 No-Free-Lunch 定理

本章では，すべての問題に対して最良の結果を出力できる学習アルゴリズムが存在しないことを示す．これを以下のように形式化し，証明することを目的とする．

定理 **1.**（*No-Free-Lunch*）ドメイン $\mathcal{X}$ 上で，$0-1$*loss* を用いた 2 値分類タスクに対する任意の学習アルゴリズムを A とする．また，m を $|\mathcal{X}|/2$ より小さい学習データ集合のサイズとする．ここで，以下を満たすような $\mathcal{X} \times \{0,1\}$ 上の分布 $\mathcal{D}$ が存在する．

1. $L_{\mathcal{D}}(f) = 0$ となるような関数 $f : \mathcal{X} \to \{0,1\}$ が存在する．
2. 少なくとも $1/7$ の確率で，$S \sim \mathcal{D}^m$ について $L_{\mathcal{D}}(A(S)) \geq 1/8$ を満たす．

この定理は，すべての学習アルゴリズムについて，必ずそのアルゴリズムが失敗するようなタスクが存在することを意味している．

Proof. C をサイズ $2m$ の $\mathcal{X}$ の部分集合とする．また，学習データで学習したアルゴリズム $A(S)$ の C，内の未知のデータに対する予測と矛盾するようなラベル付けを行う関数 f を考える．ここで，$f : C \to \{0,1\}$ のような関数は $T = 2^{2m}$ 通り存在する．それらの関数を $f_1, \ldots, f_T$ で表す．各関数に対して，$\mathcal{D}_i$ を $C \times \{0,1\}$ 上の分布とする．

$$\mathcal{D}_i(\{x,y\}) = \begin{cases} 1/|C| & if/y = f_i(x) \\ 0 & otherwise. \end{cases} \tag{1}$$

明らかに，$L_{\mathcal{D}_i}(f_i) = 0$ である．
任意のアルゴリズム A について，以下が成り立つことを証明する．

$$\max_{i\in[T]} \mathbb{E}_{S\sim D_i^m}[L_{\mathcal{D}_i}(A(S))] \geq 1/4. \tag{2}$$

これは，各アルゴリズム A' について，

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A'(S))]] \geq 1/4. \tag{3}$$

C からのサイズ m の列の取り出し方は $k = (2m)^m$ 通りの方法がある．そのような列を $S_1, \ldots, S_k$ とする．また，$S_j = (x_1, \ldots, x_m)$ のとき，関数 f_i でラベル付けされた S_j 内のインスタンス列を $S_j^i = ((x_1, f_i(x_1))), \ldots, (x_m, f_i(x_m))$ とすると，

$$\mathbb{E}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] = \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)). \tag{4}$$

最大，平均，最小の関係性から，

$$\begin{aligned}
\max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) \geq \quad & \tfrac{1}{T}\textstyle\sum_{i=1}^{T} \tfrac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)) && (5)\\
= \quad & \tfrac{1}{k}\textstyle\sum_{j=1}^{k} \tfrac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) && (6)\\
\leq \quad & \min_{j\in[k]} \tfrac{1}{T}\textstyle\sum_{i=1}^{T} L_{D_i}(A(S_j^i)). && (7)
\end{aligned}$$

次に，ある $j \in [k]$ を固定し，学習サンプルを $S_j = (x_1, \ldots, x_m)$，S_j に含まれない C のサンプルを $v_1, \ldots, v_p$ とする．ここで，$p \geq m$，関数 $h : C \to \{0, 1\}$ とすると，

$$\begin{aligned}
L_{\mathcal{D}_i}(h) = \quad & \tfrac{1}{2m}\textstyle\sum_{x\in C} \mathbb{1}_{[h(x)\neq f_i(x)]} && (8)\\
\geq \quad & \tfrac{1}{2m}\textstyle\sum_{r=1}^{p} \mathbb{1}_{[h(v_r)\neq f_i(v_r)]} && (9)\\
\geq \quad & \tfrac{1}{2p}\textstyle\sum_{r=1}^{p} \mathbb{1}_{h(v_r)\neq f_i(v_r)}. && (10)
\end{aligned}$$

ここで，

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \geq \quad & \tfrac{1}{T}\textstyle\sum_{i=1}^{T} \tfrac{1}{2p}\sum_{r=1}^{p} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} && (11)\\
= \quad & \tfrac{1}{2p}\textstyle\sum_{r=1}^{p} \tfrac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} && (12)\\
\geq \quad & \tfrac{1}{2}\cdot\min_{r\in[p]} \tfrac{1}{T}\textstyle\sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]}. && (13)
\end{aligned}$$

関数列 $f_1, \ldots, d_T$ を，$f_i(c) \neq f_j(c), c \in C$ となるような $T/2$ の異なるペアに分割することができる．このようなペアは自明に以下を満たす．

$$\mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} + \mathbb{1}_{[A(S_j^i)(v_r)\neq f_{i'}(v_r)]} = 1. \tag{14}$$

これは,

$$\frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r)\neq f_i(v_r)]} = \frac{1}{2}. \tag{15}$$

式 15，式 11，式 5，式 4から，式 2を導出できる．

□

1.1 No-Free-Lunch and Prior Knowledge

仮説集合 $\mathcal{H}$ 上の ERM 予測器を考える．この仮説集合は事前知識を欠いており，すべてのありうる予測器が良い候補だとみなされている．No-Free-Lunch 定理に則ると，任意の学習アルゴリズムは必ずいくつかのタスクで失敗するような予測器しか出力できない．したがって，このようなクラスは PAC 学習不可能である．

これを受けて，我々はある特定のタスクに対して，事前知識を活用することが必要となる．

2 Error Decomposition

$\mathrm{ERM}_{\mathcal{H}}$ について，以下のように書くことができる．

$$L_{\mathcal{D}}(h_S) = \epsilon_{app} + \epsilon_{est} \quad where: \ \epsilon_{app} = \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h), \ \epsilon_{est} = L_{\mathcal{D}}(h_S) - \epsilon_{app}. \tag{16}$$

- **The Approximation Error**：仮説集合内の予測器が達成しうる最小のリスク．この値は,我々がどれだけ inductive bias を持っているかの指標となる.Approximation error は，実現可能性を仮定する場合は 0 になる一方で，仮定しない場合は非常に大きくなってしまう．
- **The Estimation Error**：Approximation error と ERM 予測器が達成したエラーとの誤差．推定の質は，学習データ集合のサイズと仮説集合のサイズに依存する．

References

[1] shai shalev shwartz and shai ben david. *understanding machine learning: from theory to algorithms.* cambridge university press, 2014.