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

Masanari Kimura

June 13, 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 Chapter3 の内容を含みます．

1 PAC Learning

前章で，十分大きな学習サンプルが与えられた際の有限仮説集合上での ERM がおそらく確からしい仮設を出力できることを示した．本章では，より一般的に Probably Approximately Correct (PAC) Learning を定義していく．

定義 **1.**（*PAC Learnability*）ある関数 $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ が存在し，学習アルゴリズムが以下のような性質を備えている場合，仮説集合 $\mathcal{H}$ は *PAC* 学習可能であるという：

任意の $\epsilon, \delta \in (0,1)^2$ と $\mathcal{X}$ 上の分布 $\mathcal{D}$，ラベリング関数 $f : \mathcal{X} \to \{0,1\}$ について，$m \geq m_{\mathcal{H}}(\epsilon, \delta)$ のサンプルが *i.i.d.* に生成されるとする．このとき学習アルゴリズムは，少なくとも $1-\delta$ の確率で $L_{(D,f)(h)} \leq \epsilon$ であるような仮説 h を出力する．

PAC 学習可能性の定義は，すなわち，十分なサイズのサンプルを用いて学習すれば任意の確率および精度で期待損失が最小に近いような仮説を得ることができることを意味する．

定義 **2.**（*Sample Complexity*）関数 $m_{\mathcal{H}} : (0,1) \to \mathbb{N}$ は $\mathcal{H}$ の学習サンプルの複雑さを決定する．

サンプルの複雑さとは，おそらく確からしい解を得るために必要なサンプルの数を意味する．サンプルの複雑さは精度 ϵ と信頼度 δ をパラメータに取る関数として表現される．これは仮説集合 $\mathcal{H}$ の性質に依存して決定され，例えば $\mathcal{H}$ が有限仮説集合であればサンプルの複雑さは仮説集合のサイズの log に依存する．

ここで，もし $\mathcal{H}$ が PAC 学習可能であれば，PAC 学習可能性を満たす複雑性関数 $m_{\mathcal{H}}$ は無数に存在するため，$m_{\mathcal{H}}(\epsilon, \delta)$ を PAC 学習可能性を満たす最小の値とする．

系 **1.** 任意の有限仮説集合は以下のサンプル複雑性で *PAC* 学習可能である．

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil \tag{1}$$

2 より一般的な学習モデル

2.1 実現可能性の除去

実現可能性は，実際の学習タスクにおいては非常に強い仮定となる．そこで，実現可能性を仮定しないような一般化が必要となる．

（再活）実現可能性は，$\mathbb{P}_{x\sim\mathcal{D}}[h^*(x) = f(x)] = 1$ を満たす $h^* \in \mathcal{H}$ が存在することを仮定している．以下では，データのラベルを生成する分布に基づく target labeling function を用いて実現可能性を緩和する．

$\mathcal{D}$ をデータ点 $\mathcal{X}$ とラベル $\mathcal{Y}$ との結合分布とする．このような結合分布は 2 以下のような 2 つのコンポーネントから成り立っているとみなすことができラる：

- ラベルなしデータの分布 $D_{\mathcal{X}}$
- 各データ点に対するラベルの条件付き確率 $\mathcal{D}((x, y)|x)$

仮説 h の真のリスクを以下のように書き換える．

$$L_{\mathcal{D}} := \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] := \mathcal{D}(\{(x, y) : h(x) \neq y\}) \tag{2}$$

このようなリスクを最小化するような仮説 h を見つけたいが，学習アルゴリズムはデータを生成する $\mathcal{D}$ についての知識を持たず，入力として得られるのは学習データ S のみとなる．経験損失は，

$$L_S(h) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m} \tag{3}$$

目的：真のリスク $L_{\mathcal{D}}(h)$ を最小化するような仮説 h を獲得したい．

The Bayes Optimal Predictor：ある確率分布 $\mathcal{D}$ が与えられたとき，最良の予測関数は，

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & if\ \mathbb{P}[y = 1|x] \geq 1/2 \\ 0 & otherwise \end{cases} \tag{4}$$

定義 **3.**（*Agnostic PAC Learnability*）ある関数 $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ が存在し，学習アルゴリズムが以下のような性質を備えている場合，仮説集合 $\mathcal{H}$ は *agnostic PAC* 学習可能であるという：

任意の $\epsilon, \delta \in (0, 1)$ と $\mathcal{X} \times \mathcal{Y}$ 上の分布 $\mathcal{D}$ について，$m \geq m_{\mathcal{H}}(\epsilon, \delta)$ のサンプルが *i.i.d.* に生成されるとする．このとき学習アルゴリズムは，少なくとも $1 - \delta$ の確率で $L_{\mathcal{D}} \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$ であるような仮説 h を出力する．

実現可能性が仮定できる場合，agnostic PAC 学習は PAC 学習と同じ保証を提供することになるため，agnostic PAC 学習は PAC 学習の一般化であると言える．

2.2 多様なタスクへの適用

実現可能性を仮定しないような一般化の他に，適用するタスクの拡張による一般化も必要である．例えば，適用先のタスクは以下のようなものが考えられる：

- 多クラス分類
- 回帰問題
- 次元削減

幅広いタスクへの適用を可能にするため，以下のような一般化を行う．

仮説集合 $\mathcal{H}$ と何らかのドメイン Z が与えられたとき，$\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ を $\mathcal{H} \times Z$ から非負の実数に写像する任意の関数とする．このような関数を一般に損失関数と呼ぶ．

また，仮説 $h \in \mathcal{H}$ の期待損失を表現するリスク関数を定義する．

$$L_{\mathcal{D}} := \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)] \tag{5}$$

これは，分布 $\mathcal{D}$ からランダムにサンプリングされた各サンプル z についての仮説 $\langle$ の損失の期待値を意味する．同様に与えられたサンプル $S = (z_1, \ldots, z_m) \in Z^m$ に対する経験リスクを定義する．

$$L_S(h) := \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i) \tag{6}$$

最後に，一般的な損失関数に対する agnostic PAC 学習可能性を定義する．

定義 **4.** ある関数 $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ が存在し，学習アルゴリズムが以下の性質を備えている場合，仮説集合 $\mathcal{H}$ は任意のドメイン Z と任意の損失関数 $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ について *agnostic PAC* 学習可能であるという．任意の $\epsilon, \delta \in (0,1)$ と Z 上の分布 D について，$m \geq m_{\mathcal{H}}(\epsilon, \delta)$ のサンプルが *i.i.d* に生成されるとする．このとき学習アルゴリズムは，少なくとも $1 - \delta$ の確率で $L_{\mathcal{D}} \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$ であるような仮説 h を出力する．ここで，$L_{\mathcal{D}} = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$ とする．

注意 **1.**（可測性について）前述の定義で，各 $\in \mathcal{H}h$ について，損失関数 $\ell(h, \cdot) : Z \to \mathbb{R}_+$ を乱数，$L_{\mathcal{D}}$ を乱数に対する期待値とした．そのため，損失関数 $ell(h, \cdot)$ は可測でなければならない．形式的には，Z の部分集合の σ-集合台数の存在を仮定する．

References

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