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

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 Chapter2 の内容を含みます．

1 Empirical Risk Minimization

学習アルゴリズムは，未知の分布 $\mathcal{D}$ からサンプリングされた学習データセット S を入力として，予測器 $h_S : \mathcal{X} \mapsto \mathcal{Y}$ を出力する．ここでこの学習アルゴリズムは，未知の $\mathcal{D}$ について損失を最小化するような h_S を見つけることが目的となる．

学習アルゴリズムは分布 $\mathcal{D}$ 全体について観測することはできないため，真の損失を直接得ることはできない．そこで一つの解決策として，以下の経験損失を最小化することが挙げられる：

$$L_s(h) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m} \tag{1}$$

ここで $[m] = \{1,\ldots,\mathrm{m}\}$．このように現実的に得られる学習サンプルについて経験損失 $L_s(h)$ を最小化することで h の学習を行うようなフレームワークを Empirical Risk Minimization（ERM）という．

1.1 Empirical Risk Minimization with Inductive Bias

ERM は強力なフレームワークである一方，過適合（Overfitting）の問題が存在する．そこで，ERM が学習データだけでなく未知のデータに対しても良好な性能を保証できるような方法を探す必要がある．

一つの一般的な解決方法として，ERM の探索空間を制限することが挙げられる．形式的には，学習アルゴリズムは学習データを観測する前に，ある予測器集合を選択しておくことになる．この予測器集合を仮説集合（hypothesis class）と呼び，$\mathcal{H}$ で表現する．与えられた仮説集合 $\mathcal{H}$ と学習データセット S について，学習アルゴリズム $\mathrm{ERM}_{\mathcal{H}}$ は ERM のフレームワークを用いて予測器 $h : \mathcal{X} \mapsto \mathcal{Y} \in \mathcal{H}$ を選択する．

$$ERM_{\mathcal{H}} \in argmin_{h \in \mathcal{H}} L_S(h) \tag{2}$$

このような学習アルゴリズムに予測器を $\mathcal{H}$ から選択するような制限を inductive bias という．こうした $\mathrm{ERM}_{\mathcal{H}}$ は過適合しないことが保証されている．

1.2 有限仮説集合

仮説集合に対するもっとも単純な制限は，仮説集合のサイズに上界（upper bound）を設けることが考えられる．本章では，仮説集合 $\mathcal{H}$ が有限である場合，学習サンプルが十分に与えられた $\mathrm{ERM}_{\mathcal{H}}$ が過適合しないことを示す．

学習サンプル S について h_S を $\mathrm{ERM}_{\mathcal{H}}$ を適用した結果とすると，

$$h_S \in argmin_{h \in \mathcal{H}} L_S(h) \tag{3}$$

とする．以降，簡単のため以下に示す仮定を用いる．

定義 **1.**（*The Realizability Assumption*）$L_{\mathcal{D}}(h^*) = 0$ となるような $h^* \in \mathcal{H}$ が存在する．

定義 **2.**（*The i.i.d. assumption*）学習サンプルは独立かつ同一な分布 $\mathcal{D}$ に従う．

未知の分布 D からサンプリングされるデータが，学習データについて非代表的なデータである場合，分類器は正しく分類を達成することができないはずである．このため，$L_{\mathcal{D}}(h_S)$ を小さくするような学習データセットのサンプルが行える確率を考える．一般的に，非代表的なデータをサンプルする確率を δ とし，$(1-\delta)$ を予測に対する confidence parameter と呼ぶ．

加えて，予測の質に関するパラメータである accuracy parameter ϵ を導入する．このパラメータについて，$L_{\mathcal{D}}(h_S) > \epsilon$ であるようなアルゴリズムを失敗とみなし，$L_{\mathcal{D}}(h_S) < \epsilon$ であるようなアルゴリズムを近似的に正しい予測器であるとみなす．したがって，m 個のサンプルが与えられた時に学習器が失敗する確率の上界を与えたい．

$$\mathcal{D}^m(\{S|_x : L_D(h_S) > \epsilon\}) \tag{4}$$

仮に，$\mathcal{H}_B$ を間違った仮説であるとする．

$$\mathcal{H}_B = \{h \in \mathcal{H} : K_D(h) > \epsilon\} \tag{5}$$

さらに，学習データセットにはよく適合しているが，未知のサンプルに対して誤分類を誘発するようなサンプル集合を M とする．

$$M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\} \tag{6}$$

ここで，実現性仮説から，$L_{\mathcal{D}}(h_S) > \epsilon$ となるのは学習サンプルが M に含まれる時であることがわかる．

$$\{S|_x : L_{\mathcal{D}}(h_S) > \epsilon\} \subseteq M \tag{7}$$

M は以下のように書き換えることができる．

$$M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\} \tag{8}$$

したがって，

$$\mathcal{D}^m(\{S|_x : L_{\mathcal{D}}(h_S) > \epsilon\}) \leq \mathcal{D}^m(M) = \mathcal{D}^m(\cup_{h \in \mathcal{H}_B \{S|_x : L_S(h)=0\}}) \tag{9}$$

ここで，Union Bound を用いて右辺の上界を与えることができる．

補題 **1.**（*Union Bound*）任意の二つの集合 A，B と分布 $\mathcal{D}$ は以下を満たす．

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B) \tag{10}$$

式 9の右辺に一様バウンドを適用することで，

$$\mathcal{D}^m(\{S|_x : L_D(h_S) > \epsilon\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \tag{11}$$

次に，ある悪い仮説 $h \in \mathcal{H}_B$ について考える．$L_S(h) = 0$ は $\forall i, h(x_i) = y_i$ に等しい．学習データは i.i.d. にサンプルされると仮定しているので，

$$\begin{aligned} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) = &\quad \mathcal{D}^m(\{S|_x : \forall i, h(x_i = y_i\}) && (12)\\ = &\quad \textstyle\prod_{i=1}^m \mathcal{D}(\{x_i : h(x_i) = y_i\}) && (13) \end{aligned}$$

独立にサンプリングされた学習データセットの各要素について，

$$\mathcal{D}(\{x_i : h(x_i) = y_i\}) = 1 - L_D(h) \leq 1 - \epsilon \tag{14}$$

不等式 $1 - \epsilon \leq e^{-\epsilon}$ を用いて，

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \leq (1 - \epsilon)^m \leq e^{-\epsilon m} \tag{15}$$

式 11と式 15から，

$$\mathcal{D}^m(\{S|_x : L_D(h_S) > \epsilon\}) \leq |\mathcal{H}_B| e^{-\epsilon m} \leq |H| e^{-\epsilon m} \tag{16}$$

系 **1.** $\mathcal{H}$ を有限仮説集合とする．$\delta \in (0,1)$ と $\epsilon > 0$ について，m を以下が成り立つ整数とする．

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \tag{17}$$

ここで，任意の分布 $\mathcal{D}$ について，実現性仮説が成り立つ時，*i.i.d.* なサイズ m のサンプル集合 S 上で少なくとも $1-\delta$ の確率で，全ての *ERM* 仮説について以下を得る．

$$L_D(h_S) \leq \epsilon \tag{18}$$

前述の系は，十分大きな m について，有限仮説上での $\mathrm{ERM}_{\mathcal{H}}$ は（$1-\delta$ の確率で）おそらく，（たかだか ϵ のエラーで）確からしいと言える（Probably Approximately Correct）．

References

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