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

Masanari Kimura

June 13, 2019

Abstract

本資料は書籍”Understanding Machine Learning: From Theory to Algorithms”[4] の輪講資料です．本資料は該当書籍の Chapter4 の内容を含みます．

1　一様収束

本章では，一様収束を用いて，有限仮設集合が実現可能性を仮定せず，一般化した損失関数について agnostic PAC 学習可能であることを示す．

定義 1.（*ϵ-representative sample*）以下を満たすとき，学習データセット S は *ϵ-representative* であるという．

$$\forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon. \tag{1}$$

以下の補題では，学習データ集合が $(\epsilon/2)$-representative であれば，ERM は常に良い仮定を出力できることを示す．

補題 1. 学習データ集合 S が *$\epsilon/2$-representative* であると仮定する．このとき，$ERM_{\mathcal{H}}(S)$ の出力 $h_S \in argmin_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$ は以下を満たす．

$$L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon. \tag{2}$$

Proof. すべての $h \in \mathcal{H}$ について，

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2}\ L_{\mathcal{D}}(h) + \epsilon. \tag{3}$$

1 つめと 3 つめの不等式は S が $\epsilon/2$-representative の仮定から，2 つめの不等式は h_S が ERM 予測器であることから導かれる．　□

前述の補題は，ERM が agnostic PAC 学習可能であるためには，学習データ集合からの無作為抽出に対して，少なくとも $1-\delta$ の確率でそれが ϵ-representative でなければいけないことを示している．

定義 2. （*Uniform Convergence*）以下の条件を満たす関数 $m_{\mathcal{H}}^{UC}:(0,1)^2 \to \mathbb{N}$ が存在するとき，仮説集合 $\mathcal{H}$ は一様収束であるという．

任意の $\epsilon, \delta \in (0,1)$ と Z 上の確率分布 D に対して，$m \geq m_{\mathcal{H}}^{UC}(\epsilon,\delta)$ の $\mathcal{D}$ から生成されるサンプルを *i.i.d.* にとったとき，少なくとも $1-\delta$ の確率で S が *ϵ-representative* である．

関数 $m_{\mathcal{H}}^{UC}$ は一様収束を満たすためにサンプル複雑性を示す．ここでのサンプル複雑性は，少なくとも $1-\delta$ の確率でサンプルが ϵ-representative であると示すために必要なサンプルの最小の数を意味する．

系 1. 仮説集合 $\mathcal{H}$ が関数 $m_{\mathcal{H}}^{UC}$ で一様収束性をもつとき，サンプル複雑性 $m_{\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2,\delta)$ で *agnostically PAC* 学習可能である．

2 一様収束を用いた PAC 学習可能性の証明

任意の ϵ, δ が与えられるとする．まず，任意の $\mathcal{D}$ について i.i.d. にサンプリングされた $S=(z_1,\ldots,z_m)$ が少なくとも $1-\delta$ の確率で $h \in \mathcal{H}, |L_S(h)-L_D(h)|$ を満たすことを保証するサンプルサイズ m を求めたい．つまり，

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h)-L_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1-\delta. \tag{4}$$

これは，以下を示すことに等しい．

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h)-L_{\mathcal{D}}(h)| > \epsilon\}) < \delta. \tag{5}$$

上式を，以下のように書き直す．

$$\{S : \exists h \in \mathcal{H}, |L_S(h)-L_{\mathcal{D}}(h)| > \epsilon\} = \cup_{h\in\mathcal{H}}\{S : |L_S(h)-L_{\mathcal{D}}(h)| > \epsilon\}. \tag{6}$$

これに，一様収束を適用すると，以下を得られる．

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h)-L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h\in\mathcal{H}} \mathcal{D}^m(\{S : |L_S(h)-L_D(h)| > \epsilon\}). \tag{7}$$

十分大きな m について，この不等式の右辺が十分小さい値を取ることを保証したい．これは，ある仮定 h について，真のリスクと経験的リスクの誤差 $|L_S(h)-L_{\mathcal{D}}(h)|$ が小さくなることを示すことに等しい．

（再活）$L_{\mathcal{D}}(h) = \mathbb{E}_{z\in\mathcal{D}}[\ell(h,z)]$，$L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)$ である．ここで，各 z_i は $\mathcal{D}$ から i.i.d. にサンプルされ，乱数 $\ell(h,z_i)$ の期待値は $L_{\mathcal{D}}(h)$ である．

期待値の線形性から，$L_{\mathcal{D}}(h)$ は $L_S(h)$ の期待値でもある．したがって，$|L_{\mathcal{D}}(h)-L_S(h)|$ の値は $L_S(h)$ の偏差になる．以降，$L_S(h)$ がその期待値の周辺に集中することを示す必要がある．

大数の法則から，m が無限大になると，$L_S(h)$ の経験的な平均はその真の期待値に収束する．しかし，大数の法則は漸近的な結果に過ぎないため，与えられた有限のサンプルサイズにおける真のエラーと経験的エラーの誤差は求められない．

代わりに，Hoeffding の不等式を導入し，これを用いて証明を行う．

補題 2. （*Hoeffding's Inequality*）$\theta_1,\ldots,\theta_m$ を *i.i.d.* な乱数列とし，すべての i について，$\mathbb{E}[\theta_i]=\mu$ かつ $\mathbb{P}[a\le\theta_i\le b]=1$ と仮定する．ここで，すべての $\epsilon>0$ について，

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i-\mu\right|>\epsilon\right]\le 2\exp(-2m\epsilon^2/(b-a)^2) \tag{8}$$

ここでは，乱数 θ_i を $\ell(h,z_i)$, $L_S(h)=\frac{1}{m}\sum_{i=1}^{m}\theta_i$, $L_{\mathcal{D}}(h)=\mu$ とする．加えて，$\theta_i\in[0,1]$ と仮定すると，

$$\mathcal{D}^m(\{S:|L_S(h)-L_{\mathcal{D}}(h)|>\epsilon\})=\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i-\mu\right|>\epsilon\right]\le 2\exp(-2m\epsilon^2). \tag{9}$$

式 7と合わせて，

$$\begin{aligned}\mathcal{D}^m(\{S:\exists h\in\mathcal{H},|L_S(h)-L_D(h)|>\epsilon\})\le&\quad \textstyle\sum_{h\in\mathcal{H}}2\exp(-2m\epsilon^2) &(10)\\ =&\quad 2|\mathcal{H}|\exp(-2m\epsilon^2). &(11)\end{aligned}$$

最後に，$m\ge\log(2|\mathcal{H}|/\delta)/2\epsilon^2$ とすると，

$$\mathcal{D}^m(\{S:\exists h\in\mathcal{H},|L_S(h)-L_D(h)|\le\epsilon\})\le\delta. \tag{12}$$

系 2. $\mathcal{H}$ を有限仮説集合，Z をドメイン，$\ell:\mathcal{H}\times Z\to[0,1]$ を損失関数とすると，$\mathcal{H}$ は以下の複雑性で一様収束性を満たす．

$$m_{\mathcal{H}}^{UC}(\epsilon,\delta)\le\left\lceil\frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}\right\rceil. \tag{13}$$

さらに，仮説クラスは *ERM* アルゴリズムを用いて，以下の複雑性で *agnostic PAC* 学習可能である．

$$m_{\mathcal{H}}(\epsilon,\delta)\le m_{\mathcal{H}}^{UC}(\epsilon/2,\delta)\le\left\lceil\frac{2\log(2|\mathcal{H}|/\delta))}{\epsilon^2}\right\rceil. \tag{14}$$

3 Bibliographic Remarks

一様収束性を満たすような関数クラスは Glivenko-Cantelli クラスと呼ばれ，初めて一様収束性について証明してた Valery Ivanovich Glivenco と Francesco Paolo Cantelli によって命名された [3]．一様収束性と学習可能性の関係性については Vapnik の一連の研究が詳しい [1, 2]．

References

[1] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[3] Richard M Dudley, Evarist Giné, and Joel Zinn. Uniform and universal glivenko-cantelli classes. *Journal of Theoretical Probability*, 4(3):485–510, 1991.

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

Hoeffding's Inequality

$\theta_1, \dots, \theta_m$ を i.i.d. な乱数列, $\overline{\theta} = \frac{1}{m}\sum_{i=1}^{m}\theta_i$ とする.
$\mathbb{E}[\overline{\theta}] = \mu$ かつ $\mathbb{P}[a \leq \theta_i \leq b] = 1$ とすると, 任意の $\epsilon > 0$ について,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2\exp(-2m\epsilon^2/(b-a)^2).$$

証明 $X_i = \theta_i - \mathbb{E}[Z_i]$ かつ $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ とする.
指数関数の単調性とマルコフの不等式を用いて, 任意の $\lambda > 0$ と $\epsilon > 0$ について,

$$\mathbb{P}[\bar{X} \geq \epsilon] \underset{\text{monotonicity of the exp function}}{=} \mathbb{P}[e^{\lambda\bar{X}} \geq e^{\lambda\epsilon}] \underset{\text{Markov's inequality}}{\leq} e^{-\lambda\epsilon}\mathbb{E}[e^{\lambda\bar{X}}].$$

独立仮定を用いると,

$$\mathbb{E}[e^{\lambda\bar{X}}] = \mathbb{E}\left[\prod_i e^{\lambda X_i/m}\right] = \prod_i \mathbb{E}[e^{\lambda X_i/m}].$$

Hoeffding の補題から,

$$\mathbb{E}[e^{\lambda X_i/m}] \leq e^{\frac{\lambda^2(b-a)^2}{8m^2}}.$$

よって,

$$\mathbb{P}[\bar{X} \geq \epsilon] \leq e^{-\lambda\epsilon}\mathbb{E}[e^{\lambda\bar{X}}]$$
$$\to \mathbb{P}[\bar{X} \geq \epsilon] \leq e^{-\lambda\epsilon}\prod_i \mathbb{E}[e^{\lambda X_i/m}]$$
$$\to \mathbb{P}[\bar{X} \geq \epsilon] \leq e^{-\lambda\epsilon}\prod_i e^{\frac{\lambda^2(b-a)^2}{8m}} = e^{-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8m}}.$$

$\lambda = 4m\epsilon/(b-a)^2$ とおくと,

$$\mathbb{P}[\bar{X} \geq \epsilon] \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

以上で Hoeffding の不等式が得られた. Q.E.D.