数理解析研究所講究録
第1938巻 2015年 54-58

On the PAC_n learning

Takayuki Kuriyama, Kota Takeuchi

Abstract

We introduce the notion of PAC_n-learning and show that the PAC_n-learnability of a class $\mathcal{C}$ implies $VC_n(\mathcal{C}) < \infty$.

1 Introduction

The PAC-learning on a class having finite VC-dimension has been studied since Vapnic and Chervonenkis published their pioneer work on VC-classes in 1970's [3]. It is well known that the PAC-learnability coincides with the finiteness of VC-dimension appears in combinatorics, and with the dependent property discussed in model theory. One of the authors introduced VC_n-dimension which is a generalization of VC-dimension to investigate n-dependent property in a joint work with Chernikov and Palacin in [2], and it is shown that VC_n-dimension characterizes n-dependency of formulas, and vice versa, in the article.

In this article we introduce a suitable generalization of PAC-learnability and discuss about relations to VC_n-dimension. This is a partial result of a joint work with M. Kobayashi. Note that readers are assumed to know basic concepts of PAC-learning and VC-dimension.

2 PAC_n-learnability and VC_n-dimension

In this section we introduce a notion of PAC_n-learning on a class $\mathcal{C}$ of subsets of $X = \Pi_{i<n} X_i$ and prove that if $\mathcal{C}$ is PAC_n-learnable then $\mathcal{C}$ has finite VC_n-dimension.

Let X be a set and let $\mathcal{B}$ be a σ-algebra on X.

Definition 1. A probability measure μ on a σ-algebra $(X, \mathcal{B})$ is a σ-additive function $\mu_i : \mathcal{B}_i \to [0,1]$ such that $\mu(\emptyset) = 0$ and $\mu(X) = 1$.

In this article, we fix sets $X_0, ..., X_{n-1}$ and $X = \Pi_{i<n} X_i$ with σ-algebras $\mathcal{B}_i$ on X_i and the product σ-algebra $\mathcal{B} = \otimes_{i<n} B_i$ on X such that for every point $x, y \in X_i$ there are disjoint sets $A, B \in \mathcal{B}_i$ satisfying $x \in A$ and $y \in B$.

Example 2. The following are typical examples of our setting.

1. X is a Euclidean space and $\mathcal{B}$ is the class of Borel sets of X.
2. X is a countable set and $\mathcal{B}$ is the power set 2^X.

For a given point $a = (b_0, ..., b_{n-1}) \in X$, we denote the set $\bigcup_{i<n}\{(x_0, ..., x_{i-1}, b_i, x_{i+1}, ..., x_{n-1}) \in X : x_j \in X_j\}$ by $D_n(a)$. Also, for a given sequence $\bar{a} \in X^m$, we denote $\bigcup_{a\in\bar{a}} D_n(a)$ by $D_n(\bar{a})$. We'll often omit the subscript of D_n when it is clear from the context.

Definition 3. Let $\mathcal{C}$ be a subclass of $\mathcal{B}$.

1. $\mathcal{C}^n_{\text{fin}} = \{f|_{D_n(\bar{a})} : f \in \mathcal{C}, \bar{a} \in X^m, m \in \omega\}$.
2. A function $H : \mathcal{C}^n_{\text{fin}} \to \mathcal{B}$ is said to be a learning function for $\mathcal{C}$ if for every $\epsilon, \delta > 0$ there is $N_{\epsilon,\delta}$ satisfying the following: For every $m \geq N_{\epsilon,\delta}$, $f \in \mathcal{C}$, and a probability measure μ_i on $(X_i, \mathcal{B}_i)$ for $i < n$

$$\mu^m(\{\bar{a} \in X^m : \mu(H(f|_{D(\bar{a})})\Delta f) > \epsilon\}) \leq \delta.$$

   where μ is the product measure $\mu_0 \times ... \times \mu_{n-1}$.
3. Them minimum $N_{\epsilon,\delta}$ witnessing that H is a learning function is called the sample complexity of H.
4. $\mathcal{C}$ is said to be PAC_n-learnable if $\mathcal{C}$ has a learning function H for $\mathcal{C}$.

2.1 PAC_n-learnability implies the finiteness of VC_n-dimension

We first recall the definition of VC_n-dimension introduced in [2].

Definition 4. Let $\mathcal{C} \subset 2^X$ where $X = \Pi_{i<n}X_i$.

1. $A \subset X$ is said to be a box of size $m \in \omega$ if $A = \Pi_{i<n}A_i$ for some m-point sets $A_i \subset X_i$.
2. $VC_n(\mathcal{C}) = \sup\{\text{size}(A) : A \subset X \text{ is a box, } \mathcal{C}|_A = 2^A\}$ where $\mathcal{C}|_A = \{f|_A : f \in \mathcal{C}\}$.

$VC_n(\mathcal{C})$ is called the VC_n-dimension of $\mathcal{C}$.

Theorem 5. Every PAC_n-learnable class has finite VC_n-dimension.

Proof. Suppose that $\mathcal{C} \subset 2^X$ is PAC_n-learnable with VC_n-dimension $\geq d$. We show that the sample complexity $N_{\epsilon,\delta}$ must be $\geq d(1 - \sqrt[n]{2(\epsilon+\delta)})$. Then, as an immediate conclusion, we have that if the class has infinite VC_n-dimension then it is not PAC_n-learnable. First notice that, since $V_n(\mathcal{C}) \geq d$, there is a box $A = \Pi_i A_i$ of size d which is shattered by $\mathcal{C}$ (hence $|A| = d^n$). By our assumption on $\mathcal{B}_i$ we can find $B_b \in \mathcal{B}_i$ for each $b \in A_i$ such that $B_b \cap B_{b'} = \emptyset$ for every $b \neq b'$. Consider a measure μ_i on X_i such that $\mu_i(A_i) = 1$ and μ_i is uniform on A_i. (Hence for any subset $B_i \subset X_i$, $\mu_i(B_i)$ is determined as $|A_i \cap B_i|/d$.) This construction gives a measure μ on $(X, \mathcal{B})$ such that $\mu(B) = |A \cap B|/d^n$ for every $B \in \mathcal{B}$. So, in what follows, we assume $X = A$, $\mathcal{C} = 2^A$, and consider the uniform measure μ on A. If $N_{\epsilon,\delta} \geq d$ then there is nothing to prove, so let $m = N_{\epsilon,\delta} \leq d$.

Claim A. For any $f \in 2^A$, the expected value of the error occurring is bounded as follows:

$$\frac{1}{d^{nm}} \sum_{\bar{a}\in A^m} \mu(H(f|_{D(\bar{a})})\Delta f) \leq \epsilon + \delta.$$

proof of Claim A. Let $B = \{\bar{a} \in A^m : \mu\{H(f|_{D(\bar{a})}\Delta f) \geq \epsilon\}$ and $B^c = A^m \setminus B$. Then

$$\sum_{\bar{a}} \mu(H(f|_{D(\bar{a})})\Delta f) \leq \sum_{\bar{a}\in B} 1 + \sum_{\bar{a}\in B^c} \epsilon \leq d^{nm}(\delta + \epsilon).$$

□

Claim B. For some $f \in 2^A$, the expected value of the error occurring has a lower bound as follows:

$$\frac{1}{d^{nm}} \sum_{\bar{a}\in A^m} \mu(H(f|_{D(\bar{a})})\Delta f) \geq \frac{(d-m)^n}{2d^n}.$$

proof of Claim B. Let $E(f) = \frac{1}{d^{nm}} \sum_{\bar{a}\in A^m} \mu(H(f|_{D(\bar{a})})\Delta f)$. Then

$$\begin{aligned} \max_{f\in 2^A}\{E(f)\} &\geq \frac{1}{|2^A|} \sum_{f\in 2^A} E(f) \\ &\geq \min_{\bar{a}\in A^m}\{\frac{1}{|2^A|} \sum_{f\in 2^A} \mu(H(f|_{D(\bar{a})})\Delta f)\}. \end{aligned}$$

Therefore it is enough to show that

$$\sum_{f\in 2^A} \mu(H(f|_{D(\bar{a})})\Delta f) \geq \frac{(d-m)^n 2^{d^n}}{2d^n}$$

for every $\bar{a} \in A^m$. However the left-hand side of the inequation is calculated as follows: by letting $D = D(\bar{a})$, $D^c = A \setminus D$, $|D^c| = k$

$$\begin{aligned} (L.H.S) &= \frac{1}{d^n} \sum_{f\in 2^A} |H(f|_D)\Delta f| \\ &= \frac{1}{d^n} \sum_{f_0\in 2^D} \sum_{f\supset f_0} |H(f|_D)\Delta f| \\ &= \frac{1}{d^n} \sum_{f_0\in 2^D} \sum_{i=0}^{k} i\binom{k}{i} \\ &= \frac{1}{d^n} \sum_{f_0\in 2^D} \sum_{i=1}^{k} k\binom{k-1}{i-1} \\ &= \frac{1}{d^n} \sum_{f_0\in 2^D} k2^{k-1} \\ &\geq \frac{1}{d^n} 2^{|D|}\{(d-m)^n 2^{k-1}\} \\ &= \frac{(d-m)^n 2^{d^n}}{2d^n}. \end{aligned}$$

Note that the third line in the above inequations is followed from the fact that f varies over 2^A extending f_0. □

By Claim A and B, we have $(d-m)^n/2d^n \leq \epsilon + \delta$. The remains are straightforward calculation. □

3 ϵ-nets

The proof of Theorem 5 we gave in the last section is almost the same for the classical case $n = 1$. Hence we expect that the converse is also shown by the same method used in the case $n = 1$. However it does not work for general case $n > 1$ because an ϵ-net doesn't exist in general, though the existence of it is the most important part of the proof of the converse. (We would like to note that Sauer-Shelah lemma, which is the main lemma to prove the existence of ϵ-nets, has a suitable generalization to VC_n-classes (see [2] for the generalization, and [1] for the case $n = 1$).) In this section we see an example of $\mathcal{C}$ such that no domain $D_n(\bar{a})$ of samples in $\mathcal{C}^n_{\text{fin}}$ is an ϵ-net for $\mathcal{C}$.

First we recall so-called VC-theorem in the case $n = 1$. The readers can find the details in [1]. For the simplicity, we assume $(X, \mathcal{B}) \in \{(\mathbb{R}^k, \mathcal{B}_0), ([0,1]^k, \mathcal{B}_1), (\omega, 2^\omega)\}$ where $\mathcal{B}_0$ is the set of Borel sets and $\mathcal{B}_1 = \mathcal{B}_0|_{[0,1]^k}$.

Definition 6. A class $\mathcal{C} \subset \mathcal{B}$ is said to be countably separable if there is a countable subset $\mathcal{C}_0 \subset \mathcal{C}$ such that for every $f \in \mathcal{C}$ there is a sequence $f_0, f_1, \ldots \in \mathcal{C}_0$ such that the sequence convergence pointwise to f, i.e. for any $x \in X$ there is N such that if $n > N$ then $f_n(x) = f(x)$.

Definition 7. Let $\mathcal{C} \subset \mathcal{B}$ be any class and μ a probability measure on $(X, \mathcal{B})$. A subset $A \subset X$ is said to be an ϵ-net for $\mathcal{C}$ if for every $f \in \mathcal{C}$, $f \cap A \neq \emptyset$ where $\mu(f) > \epsilon$.

Proposition 8. Let $\mathcal{C} \subset \mathcal{B}$ have finite VC-dimension d. Suppose that $\mathcal{C}$ is countably separable. Then for any probability measure μ on $(X, \mathcal{B})$ and $\epsilon > 0$, $\mu^m(\{\bar{a} \in X^m : \bar{a} \text{ is not an } \epsilon\text{-net for } \mathcal{C}\}) \leq 2((2m)^d + 1)2^{-\epsilon m/2}$ where μ^m is the product measure on X^m.

By taking m large enough, we have the following:

Corollary 9. Let $\mathcal{C} \subset \mathcal{B}$ have finite VC-dimension d. Suppose that $\mathcal{C}$ is countably separable. Then there is a finite set $A \subset X$ such that A is an ϵ-net for $\mathcal{C}$.

In [], Blumer and the coauthors showed the above proposition under an weaker assumption, well-behavedness, which is implied from the countably separability. The following example, which appears in the article, shows that if we omit the assumption from the proposition then it does not hold.

Example 10. Here we assume Continuum Hypothesis. Suppose $\mathbb{R}$ has cardinality ω_1 and let $\{r_\alpha : \alpha < \omega_1\}$ be an well-ordered enumeration of $\mathbb{R}$. Let $\mathcal{C} = \{R_\alpha : \alpha < \omega_1\}$ where $R_\alpha = \{r_\beta : \alpha < \beta < \omega_1\}$. Let μ be the standard Borel measure on $\mathbb{R}$. (So the completion of μ is the Lebesgue measure on $\mathbb{R}$.) Clearly, R_α has μ-measure 1, since it is co-countable set. So, the class $\mathcal{C}$ has no finite ϵ-net for every $0 < \epsilon < 1$. Also one can easily see that $\mathcal{C}$ is not countably separable (and not well-behaved with an argument using Fubini's theorem). This example means that even though we work in ZFC we cannot omit the assumption.

Finally, for $X = [0,1]^2$ and the Borel sets $\mathcal{B}$ of X, we give a class $\mathcal{C} \subset \mathcal{B}$ such that $VC_2(\mathcal{C}) = 1$ and $\mathcal{C}$ is countably separable but it has no ϵ-net A of the form $A = D_2(\bar{a})$ where $\bar{a} \in X^m$ and $m \in \omega$.

Example 11 (Non-existence of ϵ-nets). Let I be the closed interval $[0,1]$ with the Borel measure $(\mu_0, \mathfrak{B})$ on I. Let $X = I^2$ be the product set with product measure $\mu = \mu_0^2$. Put $\mathcal{C} = \{A_0 \times A_1 : A_i$ is a finite union of open intervals in $I\}$. One can see that $\mathcal{C}$ is countably separated by considering open intervals (q, q') with $q, q' \in \mathbb{Q}$. However, for any $1 > \epsilon > 0$, $m \in \omega$ and $\bar{a} \in X^m$, there is $f \in \mathcal{C}$ such that $D_2(\bar{a}) \cap f = \emptyset$ with $\mu(f) > \epsilon$. Therefore, random samples $D_2(\bar{a})$ cannot be an ϵ-net of $\mathcal{C}$.

This example suggests us that the VC-theorem does not hold in our general setting without changing the assumption. Seemingly the sample data $D_n(\bar{a})$ is too small to be an ϵ-net or the assumption of the countably separability does not match in our setting.

References

[1] A. Blumer, et al. "Learnability and the Vapnik-Chervonenkis dimension." Journal of the ACM (JACM) 36.4 (1989): 929-965.

[2] A. Chernikov, D. Palacin and K. Takeuchi, "On n-dependence." arXiv:1411.0120

[3] Vapnik, Vladimir N., and A. Ya Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities." Theory of Probability and Its Applications 16.2 (1971): 264-280.