DNFで表現される多次元実数領域の学習について

6C－9

中村 篤祥
NEC C&C 情報研究所

1 はじめに

n 次元実数領域 $[0,1)^n$ 上の部分領域を、$[0,1)^n$ 上の一様分布に従って得られる点及びその点が部分領域の内外どちらかという情報から部分領域を学習する問題を扱う。但し、部分領域は次の条件を満たすものに制限する。($[0,1)^n$ 上の点を $(x_1, x_2, ..., x_n)$ とする。) 1) 原子論理式が $(x_i \in R)$ という形の DNF(積和標準形) の論理式で表現される。2) R は区間 $[0,1)$ を高々 k(固定) 個の領域に分割するような領域 $\bigcup_{i=1}^{l}[a_i, b_i)$ とする。3) 各 x_i は論理式の中で高々1度しか現れない。

[2] では、分布を一様分布に固定した PAC (Probably Approximately Correct) 学習モデルにおいて μDNF(各変数が高々1度しか現れない DNF) ブール式で表現されるブール関数が学習可能であることが示されている。本稿は上記 2) の制約の下で彼らの結果の定義域を $\{0,1\}^n$ から $[0,1)^n$ に拡張したものである。

2 学習モデル

まずここで採用する PAC 学習モデルの定義を述べる。
[表記]
X : 学習領域
X_n : 長さが高々 n の X の要素の集合
C : 概念クラス ($C \subset 2^X$、表現クラスと同一視)
$C_{n,s}$: X_n の要素のみからなる概念の集合で表現形の長さが高々 s のもの
また、$x \in X, c \in C$ に対し

$$in(x,c) = \begin{cases} 1 & x \in c \\ 0 & otherwise \end{cases}$$

とする。ここでは、X 上の分布は一様分布に固定する。$h \in C$ が $c \in C$ の (one-oracle モデルでの)ϵ 近似であるとは、$h \triangle c$($\triangle$ は対称差) の確率が高々 ϵ であることをいう。上の記号を使って one-oracle モデルに関する学習可能性は次のように定義される。

C が (one-oracle モデルで) 学習可能 $\Leftrightarrow$
任意の $0 < \epsilon, \delta < 1$ と $n, s \geq 1$ と任意の $c \in C_{n,s}$ に対し X_n 上の一様分布に従って $x \in X_n$ を選び例 $(x, in(x,c))$ をくれる oracle が与えられたとき、少なくとも $1-\delta$ の確率で c の ϵ 近似仮説 $h \in C$ を出力し、実行時間が $\frac{1}{\epsilon}, \frac{1}{\delta}, n, s$ の多項式で押えられるアルゴリズムが存在する。

本稿では one-oracle モデルで学習可能性を考える。two-oracle モデル (概念内、概念外で別々に分布を考え、別々の oracle により例が得られ、エラーも別々に扱うモデル) とは、distribution-free の場合には同等であることが証明されている [1]。一様分布の場合は、彼らの証明の方法をそのまま適用して two-oracle モデルの学習可能性から one-oracle モデルの学習可能性を導くことができるが、逆は明らかではない。以下で示す学習問題は one-oracle モデルで考えることにより、より簡単なアルゴリズムで学習可能性を示すことができる。

Learning Multidimensional Euclidean Resions Representable by DNF.
Atsuyoshi Nakamura.
C&C Information Technology Research Laboratories NEC Corp.

3 多次元実数領域の学習

まず、第1節の 2) の条件に関する定義を述べる。
$X = [0,1)$ とし、$\mathcal{X} = \bigcup_{k \in \mathcal{N}}\{\bigcup_{i=1}^{k}[a_i, b_i) : \forall i \leq k, 0 \leq a_i < b_i \leq 1\} \cup \{\emptyset\}$ とする。$R \in \mathcal{X}$ に対し R を構成する連続領域の数を $rn(R)$ とする。つまり、

$$rn(R) = \min\{k : \bigcup_{i=1}^{k}[a_i, b_i) = R \quad \forall i \leq k, 0 \leq a_i < b_i \leq 1\}$$

と定義する。さらに $pn(R) = rn(R) + rn(X - R)$ とし、$\mathcal{X}_k = \{R \in \mathcal{X} : pn(R) \leq k\}$ と定義する。$\mathcal{X}_k$ は、X が高々 k 個の領域に分割されるような $\mathcal{X}$ の要素全体を表す。例えば、

$$\begin{aligned} \mathcal{X}_1 &= \{\emptyset, X\} \\ \mathcal{X}_2 &= \{[0,a) : 0 < a < 1\} \cup \{[a,1) : 0 < a < 1\} \cup \mathcal{X}_1 \\ \mathcal{X}_3 &= \{[a,b) : 0 < a < b < 1\} \\ &\quad \cup \{[0,a) \cup [b,1) : 0 < a < b < 1\} \cup \mathcal{X}_2 \end{aligned}$$

となる。

X 上の変数 $x_i (i = 1, ..., n)$ と $\mathcal{X}(\mathcal{X}_k)$ の要素 R から成る述語 $(x_i \in R)$ の集合を $\mathcal{A}(\mathcal{A}_k)$ とする。$\mathcal{A}$ の要素を原子論理式とし、それらと $\wedge$ から作られる論理式を項とし、項を $\vee$ で繋いだ形の論理式 (DNF) を考える。任意の変数 x_i を含む原子論理式を高々1つしか含まないとき μDNF と呼ぶ。この場合、x_i を含む唯一の原子論理式 $(x_i \in R)$ の R を R_i で表し、R_i の長さを l_i とする。

証明で使われる確率の表記法について説明する。$Pr(A)$ は事象 A の起こる確率を表す。$Pr^-(A)$ は負例が引かれるという条件のもとでの事象 A の起こる確率を表す。全体の一様分布で考えている場合、これは負例上の一様分布に対する確率を表す。ある試行を独立に何回か行なった場合、事象 A の生起割合 (生起回数を試行回数で割ったもの) を $\widehat{Pr}(A)$ で表す。

次に主なツールとなる不等式を挙げておく。

(Chernoff bound) m 回の独立試行で、生起確率 p の事象が高々 r 回起こる確率を $LE(p, m, r)$ とすると、次の不等式が成り立つ。

$$LE(p, m, (1-\alpha)mp) \leq e^{-\frac{\alpha^2 mp}{2}}$$

(Hoeffding の不等式) m 回の独立試行で、生起確率 p の事象の生起回数を S_m とすれば次の不等式が成り立つ。

$$Pr(|\frac{S_m}{m} - p| \geq \alpha) \leq 2e^{-2\alpha^2 m}$$

定理 3.1 X^n の一様分布に従って *one-oracle* モデルで例が得られる場合、$\mathcal{A}_2$ の要素を原子論理式とする μDNF は学習可能である。

(証明) まず、学習アルゴリズムは次のようになる。

[使用する値の定義]
$k = \lceil\frac{4n}{\epsilon}\rceil$: X の分割数
$I_j = [\frac{j-1}{k}, \frac{j}{k}), j = 1, 2, ..., k$: X の分割区間
$M = n(2 + \lfloor\log_2(k-2)\rfloor) + \frac{n(n+1)}{2}$: $\widehat{Pr^-}$ の計算回数

$\delta_1 = \frac{\delta}{2M}$: $\widehat{Pr^-}$ に要求する信頼度
$\epsilon_1 = \frac{\epsilon}{20k^2n}$: $\widehat{Pr^-}$ に要求する正確度
$m_1 = \frac{1}{2\epsilon_1^2}\ln\frac{2}{\delta_1}$: $1-\delta_1$ の確率で $\widehat{Pr^-}$ が Pr^- の ϵ_1 近似になるために必要な sample 数

Step 1 $m = \max(\frac{8}{\epsilon}\ln\frac{2}{\delta}, \frac{2}{\epsilon}m_1)$ 個の例を引く。負例が m_1 個未満ならば恒等関数1を出力してSTOP。

Step 2 **(1)** $S = \emptyset$ とする。各 $1 \le i \le n$ について $\widehat{Pr^-}(x_i \in I_1)$ と $\widehat{Pr^-}(x_i \in I_k)$ を m_1 個の負例を使って求める。どちらか一方のみが $\frac{1}{k} + \frac{\epsilon}{4kn}$ 以上であればその値を y_i とし S に要素 (i, y_i) を加える。

(2) S を y_i でソートする。y_i の最小値と $\frac{1}{k} + \frac{\epsilon}{4kn}$ との差が $2\epsilon_1$ より大きければ次へ進む。$\frac{1}{k} + \frac{\epsilon}{2kn}$ から y_i の最大値を引いた値が $2\epsilon_1$ より大きければ恒等関数0を出力してSTOP。そうでなければ小さい順に隣あった二つの y_i の値の差を求め $2\epsilon_1$ より大きくなるところを探す。見つけたらそのペアの小さい方以下はすべて S から除く。

$S' = \{i : (i, y_i) \in S\}$ とする。

(3) S' の各々の要素 i に対し、$\widehat{Pr^-}(x_i \in I_{j_0})$ と $\widehat{Pr^-}(x_i \in I_{j_0+1})$ の一方のみが $\frac{1}{k} + \frac{\epsilon}{4kn}$ 以上であるような j_0 を求める。

$\widehat{Pr^-}(x_i \in I_{j_0}) > \widehat{Pr^-}(x_i \in I_{j_0+1})$ の場合

$\widehat{R_i} = [\frac{j_0-1}{k}, 1)$

$\widehat{Pr^-}(x_i \in I_{j_0}) < \widehat{Pr^-}(x_i \in I_{j_0+1})$ の場合

$\widehat{R_i} = [0, \frac{j_0+1}{k})$

とする。但し、$\widehat{R_i} = [0, 1)$ となった i は S' から除く。$\widehat{R_i}$ の長さを $\widehat{l_i}$ とする。

Step 3 S' に含まれる i, j に対し $\widehat{Pr^-}(x_i \notin \widehat{R_i})$, $\widehat{Pr^-}(x_i \notin \widehat{R_i} \wedge x_j \notin \widehat{R_j})$ を m_1 個の負例より求める。

$$\begin{aligned} &|\widehat{Pr^-}(x_i \notin \widehat{R_i} \wedge x_j \notin \widehat{R_j}) - \widehat{Pr^-}(x_i \notin \widehat{R_i})\widehat{Pr^-}(x_j \notin \widehat{R_j})| \\ \ge\ &|\widehat{Pr^-}(x_i \notin \widehat{R_i} \wedge x_j \notin \widehat{R_j}) - (1 - \widehat{l_j})\widehat{Pr^-}(x_i \notin \widehat{R_i})| \end{aligned}$$

が成り立てば同じグループに入れ、成り立たなければ違うグループに入れる。各々のグループで項を作り $\vee$ で繋げてDNFを形成する。

証明の詳細は省略し、ポイントとなる確率についてのみ述べる。まず、Step 2で推定する確率 $Pr^-(x_i \in I_j)$ は次のようになる。

x_i を含む原子論理式をもつ項がない場合

$Pr^-(x_i \in I_j) = \frac{1}{k}$

x_i を含む原子論理式をもつ項がある場合

$A : (x_i \in R_i)$ を含む項を満たす事象

$I_j \cap R_i = \emptyset$ の場合

$Pr^-(x_i \in I_j) = \frac{1}{k} + \frac{1}{k}\frac{Pr(A)}{1-Pr(A)}$

$I_j \subset R_i$ の場合

$Pr^-(x_i \in I_j) = \frac{1}{k} - \frac{1}{k}\frac{1-l_i}{l_i}\frac{Pr(A)}{1-Pr(A)}$

また、Step 3で推定する確率は次の通りである。

x_i がある項に含まれている場合の $Pr^-(x_i \notin \widehat{R_i})$

$Pr^-(x_i \notin \widehat{R_i}) = \frac{1-\widehat{l_i}}{1-Pr(A)}$

ある項に含まれている x_i, x_j に対する

$Pr^-(x_i \notin \widehat{R_i} \wedge x_j \notin \widehat{R_j})$

x_i, x_j が違う項に含まれている場合

$$Pr^-(x_i \notin \widehat{R_i} \wedge x_j \notin \widehat{R_j}) = Pr^-(x_i \notin \widehat{R_i})Pr^-(x_j \notin \widehat{R_j})$$

x_i, x_j が同じ項に含まれている場合

$$Pr^-(x_i \notin \widehat{R_i} \wedge x_j \notin \widehat{R_j}) = (1 - \widehat{l_j})Pr^-(x_i \notin \widehat{R_i})$$

必要事例数は $O(n^6(\ln n)\frac{1}{\epsilon^7}(\ln\log\frac{1}{\epsilon})\ln\frac{1}{\delta})$ であり、実行時間は明らかに $m, n, \frac{1}{\epsilon}$ の多項式で押えられる。 □

系 3.1 X^n の一様分布に従って *one-oracle* モデルで例が得られる場合、$\mathcal{A}_h(h \in \mathcal{N})$ の要素を原子論理式とする *μDNF* は学習可能である。

(証明) $\mathcal{A}_1$ を要素とする述語は恒等的に真か偽かのどちらかなので明らかである。$h \ge 2$ の場合について前定理を h の値の影響が少なくなるように注意して一般化する。

アルゴリズムの変更部分は次の通りである。

[使用する値の定義]

$k = \lceil\frac{8(h-1)n}{\epsilon}\rceil$: X の分割数

$M = kn + \frac{n(n+1)}{2}$: $\widehat{Pr^-}$ の計算回数

$\epsilon_1 = \frac{\epsilon^2}{40kn^2}$: $\widehat{Pr^-}$ に要求する正確度

Step 2 **(1)** $S = \emptyset$ とする。各 $1 \le i \le n$ について $\widehat{Pr^-}(x_i \in I_j)$, $j = 1, 2, ..., k$ を m_1 個の負例を使って求める。$\frac{1}{k} + \frac{\epsilon}{4kn}$ 以上の値 y_i が存在すれば S に要素 (i, y_i) を加える。

(3) S' の各要素 i に対し、$\widehat{R_i}$ を次のように定める。

$$\begin{aligned} Y_i &= \{j : \widehat{Pr^-}(x_i \in I_j) < \frac{1}{k} + \frac{\epsilon}{4kn}\} \\ \widehat{R_i} &= \bigcup_{\{j-1,j,j+1\} \cap Y_i \neq \emptyset} I_j \end{aligned}$$

但し $\widehat{R_i}$ の長さが $1 - \frac{\epsilon}{4n}$ を越えた i は S' から除く。

明らかであるので細かい証明は省略する。

必要事例数は $O(h^2(\ln h)n^6(\ln n)\frac{1}{\epsilon^7}(\ln\frac{1}{\epsilon})\ln\frac{1}{\delta})$ となる。□

上の系は学習領域 X^n を次のようにした場合でも成り立つ。

- $X = [A, B), A, B \in \mathcal{R}$
- $X = (A, B]$ または $X = (A, B)$ または $X = [A, B]$
- $X = \{x : x$ は円周 C 上の点$\}$。このとき k が奇数のときには $\mathcal{X}_k = \emptyset$ となる。
- 全空間を $X_1 \times X_2 \times ... \times X_n$ とし X_i が上に示した X として許される領域のいずれかであるもの。

4 おわりに

一様分布上での μDNF の学習における [2] の結果の簡単な拡張について考察した。定義域を $\{0,1\}^n$ から $[0,1)^n$ に拡張し、DNFの形のみならず各変数の (制限された形の) 領域まで推定する問題でも一様分布において PAC学習可能であることを示した。

謝辞

日頃貴重な助言を頂き、たいへんお世話になっている安倍主任に感謝致します。

参考文献

[1] D. Haussler , M. Kearns , N. Littlestone , M. K. Warmuth . Equivalence of Models for Polynomial Learnability . *Information and Computation*, 95 (2) , 1991 , pp. 129-161 .

[2] M. Kearns , M. Li , L. Pitt , L. G. Valiant . On the learnability of Boolean formulae . *Proc. of the 19th ACM Symp. on Theory of Computing* , 1987 , pp. 285-295 .