単調な関数をふくむ確率的規則の学習について

4E−10

竹内純一、大野和彦、安倍直樹、山西健司
日本電気株式会社　C&C情報研究所

1．はじめに

連続値データに関する確率的規則の学習問題について、一つの仮説空間を提案し、その学習可能性を示し、必要な事例数の上界を求めた。

確率的規則の学習問題は、Kearns & Schapire の確率的概念の学習([KS 90])、Abe & Warmuth による確率オートマトンの学習([AW 90])（ただし、これは確率密度の学習である）、Yamanishi による、確率的規則の学習([Yam 90])などが知られている。これらはいずれも Valiant の PACモデル([Val 84])の拡張である。本稿で論ずるのは、 Haussler によって実数値関数の頑健な学習モデルに拡張された PAC モデル([Hau 89])を、確率的規則の学習問題に適用したものである。この意味で本研究は、前者二つ、特に Kearns & Shapire の研究と関連が深い。また、仮説の評価基準には、様々な距離関数が使われるが（例えば、quadratic distance[KS 90], Hellinger distance,Variation distance[Yam 90], KL divergence[AW 90]。また、[Yam 91]参照。）、ここでは特に、quadratic distance に限る。

提案する仮説空間は、確率的決定リストの前提部に、実数体を領域とし確率的に真偽が決まる述語を用いたものである。決定リストは Rivest([Riv 87]) によって提案され、Yamanishi([Yam 90]) および Kearns & Schapire ([KS 90])により確率的に拡張された。本稿で提案する仮説空間は、これらの連続値領域への一拡張である。

本稿ではこの仮説空間を、天候予測問題を例に説明し、最も単純なものに制限した場合に多項式時間で頑健に学習可能であることを示す。

2．仮説空間

提案する仮説空間は、例えば次のような決定リストからなる。

```
雨 <- A          ;p1
雨 <- B          ;p2
雨 <-            ;p3        例1
```

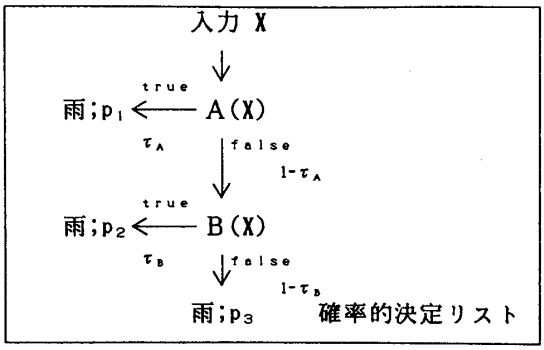
ここで前提部は、「湿度(X)∧高い(X)∧風向(東)」などのリテラルの連言である。それぞれのアトムは定数が入力されても真偽は一意には定まらず、確率的に割り付けられる。すなわち、前提部は変数の値に対する条件付き確率によって真偽が決まる。

各アトムの真偽は独立な確率変数であるとし、真となる確率を真理値と呼び τ_A などで表す。独立であるから、連言の真理値は積で計算される。また、¬A の真理値は $1-\tau_A$ である。

例１において、「雨」が真となる確率を f とし、A,B が同一のアトムを含まないと仮定すると、

$$f=p_1\tau_A+p_2(1-\tau_A)\tau_B+p_3(1-\tau_A)(1-\tau_B)$$

となる。A,B が同一のアトムを含む場合は、論理的に煩雑な操作を要する。これを避けるために、同一のアトムも、異なる位置に書かれていれば別な確率変数とみなすことにする。（この立場は確率論理([大野 89])で採用されている。）



述語には pos(X;α,μ) を用いる。 pos(·) の真理値は X の関数として、次のロジスティック関数で定義する。

$$\tau(\mathrm{pos}(X;\alpha,\mu))\equiv(1+\exp(-\alpha(X-\mu))^{-1}$$

これは、上に出てきた述語の例でいえば

高い(X)⇆pos(X;α,μ)

等に用いる。このほか決定的な述語「風向(東)」なども用いる。次に、上に述べた形式をより正確に定式化する。

［定義１］　a-アトム

a-アトム≡$\mathrm{pos}_i(X;\alpha_i,\mu_i)$

ここで$-1\leqq X\leqq 1,-a\leqq\alpha_i\leqq a,-1\leqq\mu_i\leqq 1$

$\tau(\mathrm{pos}_i(X;\alpha_i,\mu_i))=(1+\exp(-\alpha_i(X-\mu_i))^{-1}$

［定義２］　**仮説空間**

仮説空間 H(n,j,a,k) の要素 h を以下のものとする。

h: $[-1,1]^n$->[0,1]

ただし h(X) の値は、以下のような確率的決定リストによって定まる、 P が真となる確率とする。

```
P<-A1  ;p1
P<-A2  ;p2
  ⋮
P<-    ;pr
```

一変数に対して高々 j 種類の a-アトムを使う。また、各 A_i は高々 k 個のリテラルを含むとする。

このとき、仮説空間の大きさについて、

$\log(|H(n,j,a,k)|)=O((nj)^k k\log(nj))$

最も複雑な仮説のパラメータ数$=O((nj)^k)$

が成り立つ。

また、ここでは確率的述語のみが現れたが、α を十分大きくとれば決定的述語を近似できる。

最も単純な仮説を具体的に書くと、

```
P<-pos(X;α,μ)   ;p1
P<-             ;p2
```

Learning stochastic rules built on probabilistic predicates. Jun-ichi Takeuchi,Kazuhiko Ohno,Naoki Abe,Kenji Yamanishi.C&C Information Technology Laboratories,NEC Corp.

となる。ここで、与えられた X に対して P が真となる確率を $h(X;\alpha,\mu)$ とすると、

$$h(X;\alpha,\mu)=p_1 g(X;\alpha,\mu)+p_2(1-g(X;\alpha,\mu))$$
$$=(p_1-p_2)g(X;\alpha,\mu)+p_2$$
$$g(X;\alpha,\mu)\equiv(1+\exp(-\alpha(X-\mu))^{-1}$$

となる。

3．学習可能性

ここでは、仮説空間 H(1,1,a,1) が PAC(Probably Aproximately Correct)の意味で頑健に学習可能であることを示す。

［定義３］　**学習アルゴリズム**

学習事例 (X_i,T_i) が $[0,1]^n\times\{0,1\}$ 上の未知の分布 D(X,T) から独立に発生しているとする。ここで X は定義２で現れたものと同じである。T は、定義２における P の真偽を表す確率変数とする。学習アルゴリズムとは、$\varepsilon>0$、$\delta\in(0,1)$、および事例を有限個受け取って仮説 h を出力するアルゴリズムである。

［定義４］　**二乗損失(quadratic loss)**

$$LOSS(T,h(X))\equiv(T-h(X))^2$$

を二乗損失関数(quadratic loss function)と呼ぶ。これを用いて計算される、

$$l(h)\equiv E_D(LOSS(T,h(X)))$$

を真の損失と呼ぶ。(E_D は D で平均をとることを意味する。) また、m 個の事例 $(X_i,T_i)(1\leqq i\leqq m)$ が与えられているとして、

$$l_{emp}(h,m)=\frac{1}{m}\sum_{i=1}^{m}LOSS(T_i,h(X_i))$$

を経験的損失と呼ぶ。

いまh_{opt}を

$$h_{opt}\equiv\arg\min\{l(h):h\in H\}$$

で定義する。

このとき以下の定理を得る。証明技法は[AW 90]の手法を二乗損失の場合に適用したものである。

［定理］

任意の D(X,T)、$\varepsilon>0$、$\delta\in(0,1)$に対し、m 個の事例、ただし、

$$m\geqq\frac{16}{\varepsilon^2}\left\{log(2^8 3^6)+4log\left(\frac{1}{\varepsilon}\right)+2log a+log\left(\frac{1}{\delta}\right)\right\}$$

をもらい、少なくとも $1-\delta$ の確率で、

$$l(h)-l(h_{opt})\leqq\varepsilon$$

を満たす $h\in H(1,1,a,1)$ を $1/\varepsilon$, $1/\delta$, a の多項式時間で出力する学習アルゴリズムが存在する。

(証明のあらまし)

仮説空間 H(1,1,a,1) (以下 H と書く) の要素は

$$h(X;\alpha,\mu)=(p_1-p_2)g(X;\alpha,\mu)+p_2$$
$$-1\leqq\mu\leqq1,\ -a\leqq\alpha\leqq a$$

である。

H の部分集合として、$H_f(\zeta,\eta,\theta)$ を構成する。これは、p,α,μ をそれぞれ ζ,η,θ の幅で量子化した空間である ([AW 90]の手法)。

このとき、

$$\exists h_f\in H_f(\zeta,\eta,\theta)$$
$$(l(h_f)-l(h_{opt}))\leqq4\zeta+(9/4)\eta+2a\theta \qquad ①$$

が成り立つ。

そこで、次のアルゴリズムを使う。まず与えられた ε に対して $H_f(\varepsilon/24,2\varepsilon/27,\varepsilon/12a)$ を構成する。このとき $|H_f|$ は $1/\varepsilon$ の多項式オーダーである。①より、$l(h_f)-l(h_{opt})\leqq\varepsilon/2$ なる h_fが存在する。上に示した事例数をもらったあと、アルゴリズムは H_f の中で α,μ をまず固定し、l_{emp} を p_1,p_2 の関数として求める。次にこれを p_1,p_2 について微分して 0 とおき p_1,p_2 について解く。これは二元一次の方程式となり、容易に解ける ([KS 90]Th.10 参照)。得られた解をまるめて、それぞれの α,μ について $l_{emp}(h_f,m)$ が最小となる $h_f^*(\alpha,\mu)$ を求める。アルゴリズムは α,μ についての全解探索により、$l_{emp}(h_f^*(\alpha,\mu),m)$ を最小とする α^*,μ^* を求め、$h^*=h_f^*(\alpha^*,\mu^*)$ を出力する。(これは多項式時間で可能である。)

有限個の確率変数 $LOSS(T,h_f(X))$ $(h_f\in H_f)$ の集合について Hoeffding不等式を用いると、上に示した事例数で、$l(h_f)$ と $l_{emp}(h_f,m)$ の差が少なくとも $1-\delta$ の確率で $\varepsilon/4$ で抑えられることが分かる。これから、

$$l(h^*)-l(h_{opt})\leqq\varepsilon$$

を得る。 □

4．おわりに

今後はH(n,j,a,k)の学習可能性を考察するとともに、ヒューリスティクスによる高速学習アルゴリズムの開発を目指していく。

さらに、現実の気象データに適用し実験を行う予定である。

5．参考文献

[AW 90] Abe,N. & Warmuth,M.(1990). On the computational complexity of approximating distributions by probabilistic automata.Proceedings of COLT '90.52-66.

[Hau 89] Haussler,D.(1989).Generalizinig the PAC model for neural net & other learning applications. Technical Report UCSC CRL-89-30, Univ. of California at Santa Cruz.

[KS 90] Kearns,M. & Shapire,R.(1990). Efficient distribution-free learning of probabilistic concepts. Proceedings of FOCS '90, 382-391.

[Riv 87] Rivest,R,L.(1987).Learning decision lists. Machine Learning,2,229-246.

[Val 84] Valiant,L.G.(1984).A theory of the learnable. Communications of the ACM,27, 1134-1142.

[Yam 90] Yamanishi,K.(1990).A Learning criterion for stochastic rules. Proceedings of COLT '90.67-81.

[Yam 91] Yamanishi,K.(1991).On the sample complexity of robust learning stochastic rules.SITA '91.To appear.

[大野 89] 大野和彦.(1989).確率論理:帰納推論と仮説の評価基準について.第三回人工知能学会全国大会論文集.61-64.