数理解析研究所講究録
第1616巻 2008年 1-15

統計的手法を用いた学習機械の解析

京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University
池田 和司
Kazushi IKEDA

1 はじめに

ニューラルネットワークはもともと脳のモデルとして考案され，単純な機能を持つ神経素子（ニューロン）を組み合わせによりどのような情報処理が可能になるかを議論してきた．各ニューロンは，入力の重み付き線形和を計算し，それを非線形変換したものを出力するに過ぎない．この非線形変換に符号関数を用いると，これは線形二分機械とみなすことができる．

線形二分機械はパーセプトロンと呼ばれ，単純なアルゴリズムにより与えられたデータを学習可能であることが示されたことから，第1次ニューロブームの契機となった．理論解析が進み，その限界が明らかになるにつれてブームは去ったが，符号関数の代わりにシグモイド関数を用いて連続関数とし，最急降下法を用いるバックプロパゲーションアルゴリズムが提案されたことから第2次ニューロブームを巻き起こした．

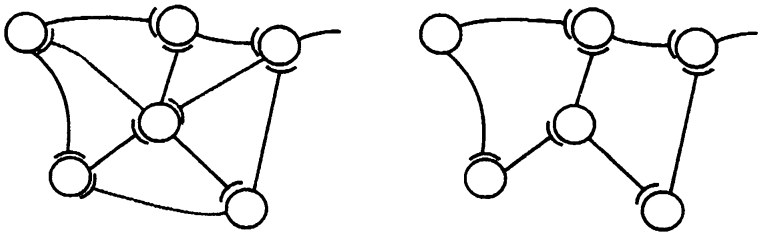
その後，研究対象は多層パーセプトロンからRBFネットワーク，混合エキスパートなどへと変わったが，学習機械の研究は脈々と続いており，近年ではサポートベクトルマシン，ブースティングなどが注目を集めている．

本講では，これらの学習機械の特性解析において，統計的な手法がどのように用いられているかを概説する．

2 柔軟な情報処理

人間はゆがんだ文字や数字でも正しく認識できる．また，提示された例題を元に，ルールを推定し未知の問題を解決する能力も高い．これと同じような機能を機械に持たせることを考えよう．例えば，手書き数字認識において，‘0’ あるいは ‘1’ という各数字についてその構成を明示的に定義し，かつその歪み具合を数値的に計る尺度を与えるのは困難である．そこで，多数の例を判別機械に学習させ，その汎化能力で未知問題をも解決させることが考えられる．

学習するシステムとして最も身近なものは生体であり，その情報処理は主として脳が担っている．したがって，学習機械のモデルとして脳を手本にしようと考えるのは自然なことである．脳は神経細胞による相互結合型ネットワーク (図 2a) であることはよく知られているが，一般にフィードバックを持つネットワークの挙動の解析は困難である．そこで，入出力の関係がより簡潔である，階層型ネットワーク (図 2b) が学習機械として用いられる．しかし，ネットワークとしての情



a) 相互結合型ネットワーク　b) 階層型ネットワーク

図 1: 神経回路網

報処理能力を考察するには，そもそも個々の要素でどのような情報処理が可能であるかを明らかにしておく必要がある．そこでまずはじめに，単一の神経細胞に着目しよう．

3　パーセプトロンの分離能力

最も簡単な神経細胞のモデルは，McCulloch-Pitts [11] によるものであり，N 次元の実数ベクトル $(x_1, x_2, \ldots, x_N) \in \boldsymbol{R}^N$ の入力に対し，

$$y = \mathrm{sgn}\left[\sum_{n=1}^{N} w_n x_n - h\right] = \mathrm{sgn}\left[\boldsymbol{w}'\boldsymbol{x}\right]$$

にしたがって $+1$ または -1 の y を出力する．ここで $\mathrm{sgn}\,[\cdot]$ は $\cdot$ が 0 以上ならば $+1$, そうでなければ -1 を出力する符号関数である．このモデルは，N 次元入力空間において，ある超平面の片側に属する入力に対して $+1$, それ以外に対して -1 を出力することから，線形二分機械と呼ばれる．また，その学習アルゴリズムにちなんでパーセプトロン (Perceptron) とも呼ばれる．

さて，パーセプトロンの学習能力はどれくらいだろうか．与えられた例題が線形分離可能ならば，これらを正しく分離する超平面は存在する，すなわち学習可能である．一方，線形分離不可能な例題は正しく分離できない．したがって，学習能力は「何個の例題まで分離できるか」で評価することができる．例えば次元 $N = 2$ で入力が T 点ある時，出力の組み合わせは 2^T 通りあるが，T 個の入力は重み空間を $2T$ 個に分割するので，$T > 2$ で $2^T > 2T$ となり，ほとんど分離でき

ないことがわかる．このような“自明な容量”ではなく，より現実的な記憶容量を算出するため，確率的評価を導入することができる．

確率的評価は，いかのように導入する．すなわち，T 点の入力に対する出力 $\{\pm 1\}^T$ が等確率で選ばれると仮定し，その例題が線形分離可能である確率 $P(T,N)$ を評価する．ここで，確率 1 で線形分離可能になる T の上限をパーセプトロンの記憶容量と定める．この時，以下の定理が成り立つ．

定理 1 入力の次元が N の時，$T < 2N$ ならば $N \to \infty$ において確率 *1* で線形分離可能．

上の定理を証明するには，T 個の入力が重み空間を何個に分割するのかを評価すればよい．

N 次元の入力が t 点与えられた時の重み空間の分割数を $C(t,N)$ としよう．新たに加えられた例題によって増える分割数が，1 だけ次元の低い超球面の分割数と一致することに注意すると (図 2)，

$$C(t+1,N) = C(t,N) + C(t,N-1)$$

という漸化式が成り立つことがわかる．したがって，ランダムに選んだ入出力関

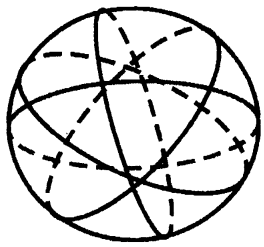


図 2: 重み空間の分割．

係が線形分離可能である確率 $P(T,N)$ は

$$P(T,N) = \frac{C(T,N)}{2^T} = \frac{1}{2^{T-1}} \sum_{n=0}^{N-1} \binom{T-1}{n}$$

と表され，$T \to \infty$ の時，$N > T/2$ ならば $P(T,N) \to 1$, $N < T/2$ ならば $P(T,N) \to 0$ である．すなわち，パーセプトロンの分離能力は $T = 2N$ であるといえる．

4 パーセプトロンの汎化能力

パーセプトロンには，収束定理および有界定理が成り立つなど，興味深い性質が数多くあるが [12], ここでは汎化能力に着目する．汎化能力とは，未学習入力を正

しく判別する能力のことであり，その評価方法は大別して二つある．一方は PAC (probably approximately correct) 学習 [16] の枠組みであり，例題数 T の時，汎化誤差が ε 以下になる確率を $1-\delta$ とし，これら 3 変数の関係を議論する．もう一方は平均汎化誤差であり，こちらは汎化誤差の例題に関する平均を評価する．本稿では後者のみを扱う．

一般に，平均汎化誤差の評価は難しく，パーセプトロンの平均汎化誤差も未解決問題である．そこでここでは，よりタイトなバウンドを求めるための手法を紹介する．単位超球面上から独立に一様に選ばれた入力と，未知の真の重みベクトル $\boldsymbol{w}_0$ を用いて T 個の例題を作成する．例題の集合を D_T と表す．学習機械がこれを学習した時，新規入力に対して正しく分離する確率を汎化誤差と定義する．求めたいものは，汎化誤差の例題に関する平均である．

ここで，例題空間を定義する．$(\boldsymbol{x}_t, +1)$ という例題と $(-\boldsymbol{x}_t, -1)$ という例題は完全に等価なので，以下では例題の出力は常に $+1$ であるとする．したがって，例題の入力 $\boldsymbol{x}_t$ は $\boldsymbol{w}_0$ との内積が非負である半球面から一様に選ばれることになる (図 3).

例題は $\boldsymbol{w}_0$ を用いて作成されているので，重み空間上で，すべての例題を正しく判別する重みの集合が定義できる．この集合を許容領域と呼び，A_T で表す．すなわち，

$$A_T = A(D_T) = \{\boldsymbol{w} | \boldsymbol{w}'\boldsymbol{x}_t > 0, t = 1, \ldots, T.\}$$

である (図 3).

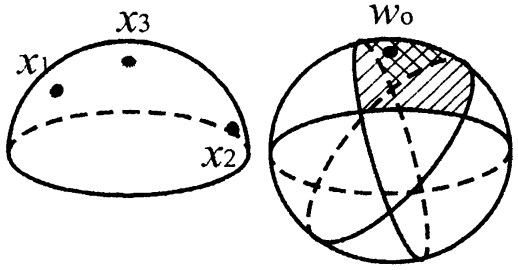


図 3: 例題空間と重み空間.

与えられた例題のうち，いくつかはなくても許容領域 A_T に影響を与えず，いくつかは削除すると許容領域が変化する．後者を有効例題と呼ぶ．有効例題は，入力空間において例題の作る凸包の頂点であることが，容易に確かめられる．

定義から明らかなように，すべての例題を正しく判別する重みは複数ある．したがって，重みの推定値 $\hat{\boldsymbol{w}}$ あるいはテスト入力 $\boldsymbol{x}_{T+1}$ の予測出力を選ぶ方法を定める必要がある．これをアルゴリズムと呼ぶ．以下ではいくつかのアルゴリズムを紹介する．

Gibbs アルゴリズム 許容領域 A_T からランダムに $\hat{\boldsymbol{w}}$ を選択する．重みの事前分布が重み空間で一様とすれば，これは事後分布にしたがって $\hat{\boldsymbol{w}}$ を選ぶことに相当する．

Bayes アルゴリズム 許容領域に属する重み $\boldsymbol{w} \in A_T$ でテスト入力 $\boldsymbol{x}_{T+1}$ に対する出力を多数決する (図 4)．これはベイズ推定に相当し，平均汎化誤差を最小にするアルゴリズムである．

最悪アルゴリズム テスト入力 $\boldsymbol{x}_{T+1}$ に対し，許容領域内のすべての重み $\boldsymbol{w} \in A_T$ で出力が一致する場合だけ正解し，そうでない場合には誤るとする．これは許容領域から重みを選ぶアルゴリズムの中で，平均汎化誤差を最大にするアルゴリズムである．

SVM マージンを最大化する重みを $\hat{\boldsymbol{w}}$ とするアルゴリズムであり，入力の大きさが同一であり，分離超平面が斉次である場合には，$\hat{\boldsymbol{w}}$ は A_T の最大内接球の中心と一致する．

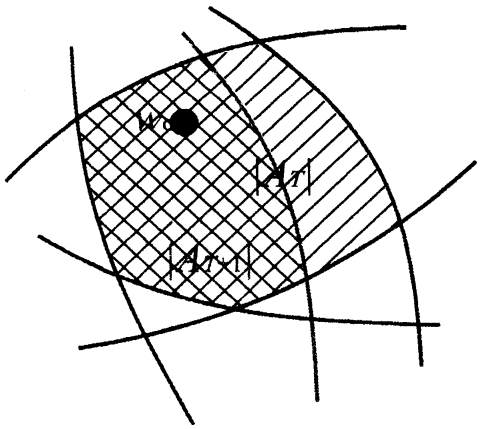


図 4: 許容領域とアルゴリズム

各アルゴリズムにおける平均汎化誤差は，許容領域の体積 $|A_T|, |A_{T+1}|$ を用い，以下のように表される．

$$\text{Gibbs アルゴリズム} \qquad \langle \varepsilon_{\mathrm{G}}(D_T) \rangle = \left\langle 1 - \frac{|A_{T+1}|}{|A_T|} \right\rangle ,$$

$$\text{Bayes アルゴリズム} \qquad \langle \varepsilon_{\mathrm{B}}(D_T) \rangle = \left\langle \mathrm{Heaviside}\left(\frac{1}{2} - \frac{|A_{T+1}|}{|A_T|} \right) \right\rangle .$$

ただし $\langle \cdot \rangle$ は例題およびテスト入力に関する期待値である．いずれの場合も，平均汎化誤差の評価には確率変数の比の期待値を含むため，その計算は困難である．なお，$Z_T = |A_T|$ と表すことにする．これは統計物理における分配関数に相当する．

Z_T の性質および対数の凸性 $1-x \leq -\log x$ を用いると，Gibbs アルゴリズムの平均汎化誤差の上限が与えられる [2]. すなわち，

$$\langle \varepsilon_{\rm G}(D_T) \rangle = \left\langle 1 - \frac{Z_{T+1}}{Z_T} \right\rangle \leq \langle \log Z_T \rangle - \langle \log Z_{T+1} \rangle$$

であるが，$Y_T = T^{N+1} Z_T$ はある Y に分布収束することから

$$\langle \log Z_T \rangle = \langle \log T^{-(N+1)} Y_T \rangle \to \langle \log Y_T \rangle - (N+1) \log T$$
$$\langle \varepsilon_{\rm G}(D_T) \rangle \leq \frac{N+1}{T}$$

が成立する.

5 積分幾何学と平均汎化誤差

本節では，積分幾何学を用いて最悪アルゴリズムの平均汎化誤差を求めてみよう [9]. 最悪アルゴリズムでは，テスト入力 $\boldsymbol{x}_{T+1}$ が許容領域と交わる時，予測は必ず誤りであると想定する．テスト入力は T 個の例題の入力と同じ分布から選ぶことから，最初から $T+1$ 個の入力を選び，そこからランダムに選んだ一つをテスト入力としても同じである．ここで，有効例題からテスト入力が選ばれる確率が [有効例題数]/$(T+1)$ であることに注意すると，平均予測誤差を求めるには平均有効例題数を求めればよいことがわかる.

前述のように，許容領域は例題空間における例題の凸包に対応している (図 5).

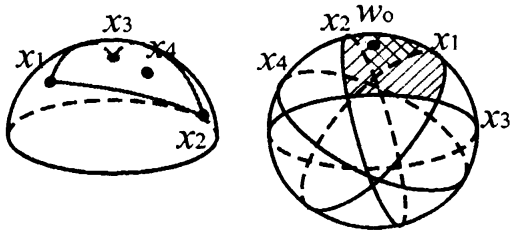


図 5: 凸包と許容領域

したがって，平均有効例題数を求めるには，例題の凸包の平均頂点数を求めればよい．T 個のランダム点群が作る凸包の統計的性質は，積分幾何学あるいは確率幾何学と呼ばれる分野の問題であり，円あるいは正方形のランダム点群の凸包の頂点数の期待値などが導出されている [4]. これを球面上の場合に応用しよう.

ある2点が作る直線 (ここでは大円) L の片側に他の $T-2$ 個の点があれば，直線 L は凸包の辺になる．したがって，直線 L が凸包の辺になる確率は

$$\left(\frac{\theta}{\pi}\right)^{N-2} + \left(1-\frac{\theta}{\pi}\right)^{N-2}$$

と表される．ある2点が作る直線 L の分布は一様分布するので，その期待値は $T \to \infty$ で漸近的に $\dfrac{\pi^2}{T(T-1)}$ となることが示される．

T 個の点から選んだ2点が作る直線の総数は ${}_T\mathrm{C}_2 = \dfrac{T(T-1)}{2}$ であるので，辺の数の平均 (= 平均頂点数) は

$$\frac{\pi^2}{T(T-1)} \cdot \frac{T(T-1)}{2} = \frac{\pi^2}{2}$$

となる．すなわち，$N=2$ の時の平均予測誤差は漸近的に $\dfrac{\pi^2}{2T}$ である．

6 多層パーセプトロン

パーセプトロンは線形分離可能な問題しか分離できない．そこで，パーセプトロンを多層化することが提案されてきた．これは多層パーセプトロン (Multi-layer Perceptron, MLP) と呼ばれる (図 6). 最も単純な MLP は，最初の層での結合をランダムに生成して固定することである．乱数の直交性および大数の法則により，次の層においてベクトルは確率的には直交し，与えられた例題は線形分離可能となる．しかし，入力を無相関にすることは，汎化能力の低下を意味する．

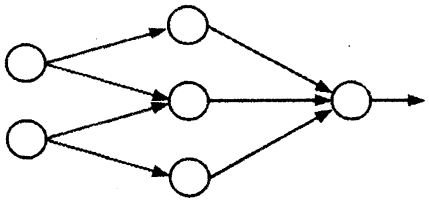


図 6: 多層パーセプトロン (MLP).

そもそも，パーセプトロンの学習曲線導出の困難さは，出力を二値とするため，不連続な符号関数を用いているところにある．そこで符号関数の代わりにシグモイド関数

$$f(x) = \frac{1}{1+\exp(-x)}$$

を用いて連続化した上で，多層化することを考えよう．この場合，MLP は重みに関して連続であるので，学習に確率降下法などをもちいることができる [1]. この

手法は Rumelhart らによって再発見され [14], 誤差逆伝播法として広く紹介され, また数値実験においても良好な成績を示したことから, 第 2 次ニューロブームを巻き起こした. また, 多層パーセプトロンが, 中間層の素子数を十分に多くとれば, 任意の連続関数を任意の精度で近似可能であり, 必要な素子数も入力の次元に依存しないことが示されたことから, 万能学習機械として注目を浴びた.

誤差逆伝播法は, 損失関数 (二乗誤差) を確率降下法で最小化するものである. すなわち

$$E(\boldsymbol{w}) = \sum_t \|z_t - y(\boldsymbol{x}_t; \boldsymbol{w})\|^2,$$
$$\Delta \boldsymbol{w} = \nabla E(\boldsymbol{w})$$

と表される. 一般に $E(\boldsymbol{x})$ は凸関数とは限らないことから局所解が多数存在するため, 実用上は, 良好な解が得られるまで多くの初期値から試す必要がある. 上式は, データにガウスノイズが付加されている, すなわち $z_t \sim N(y(\boldsymbol{x}_t; \boldsymbol{w}_o), \Sigma)$ と考えれば, $\boldsymbol{w}$ は単なる最尤推定量とみなすことができる. そこで, MLP を含む学習機械の漸近的な平均汎化誤差を導出しよう [13]. 最尤推定なので勾配がゼロであり, テイラー近似すると

$$\begin{aligned} 0 &= \frac{1}{T}\sum_{t=1}^{T} \nabla \log p(\boldsymbol{x}_t; \hat{\boldsymbol{w}}) \\ &\approx \frac{1}{T}\sum_{t=1}^{T} \nabla \log p(\boldsymbol{x}_t; \boldsymbol{w}_o) + \frac{1}{T}\sum_{t=1}^{T} \nabla^2 \log p(\boldsymbol{x}_t; \boldsymbol{w}_o)(\hat{\boldsymbol{w}} - \boldsymbol{w}_o) \end{aligned}$$

が成り立つ. ここで, 大数の法則および中心極限定理により

$$\frac{1}{T}\sum_{t=1}^{T} \nabla^2 \log p(\boldsymbol{x}_t; \boldsymbol{w}_o) \to -G$$
$$\frac{1}{T}\sum_{t=1}^{T} \nabla \log p(\boldsymbol{x}_t; \boldsymbol{w}_o) \sim N\left(0, \frac{1}{\sqrt{T}}G\right).$$

と近似すれば,

$$\langle \log p(\boldsymbol{x}, \hat{\boldsymbol{w}}) \rangle = \langle \log p(\boldsymbol{x}, \boldsymbol{w}_o) \rangle + \frac{M}{T}$$

が得られる. ここで M はパラメータ数である.

7 特異モデルの学習理論

多層パーセプトロン (MLP) のもう一つの特徴は, 確率降下法の過程において, 学習誤差がなかなか減らない減少が見られることである. これはプラトーと呼ば

れ，MLPが階層的なモデルであることがその一因であることが指摘されている[5]. すなわち，階層モデルは異なるパラメータで同一の入出力関係を表すことができる特異性を持つため，その挙動は通常の正則なモデルとは異なる．例えば，前述のように正則なモデルでは平均汎化誤差は一般にパラメータ数 M に比例するが，特異モデルでは行列 G が正則とはならないため，上の議論は成り立たない．実際，特異モデルにおけるBayesアルゴリズムについては，自由エネルギーが例題数 T，モデルで定まる有理数 λ_1, m_1 を用いて

$$F(T) = \lambda_1 \log T - (m_1 - 1) \log \log T + C$$

と表されることが知られている．ただし定数は，正則モデルの時は $2\lambda_1 = M, m_1 = 1$ であり，特異モデルの時は $2\lambda_1 < M, m_1 \geq 1$ である[18]. 平均汎化誤差は自由エネルギーの差 $F(T+1) - F(T)$ で表されるので，これは特異モデルは正則モデルよりも小さな平均汎化誤差を持つことを意味している．

8 カーネル法による多層化

近年注目を浴びているサポートベクトルマシン(Support Vector Machine, SVM)は，二乗誤差の代わりにマージン最大化を用い，多層化にカーネル法を用いた手法である[17]. 本節では，カーネル法の性質について概説する．

カーネル法における中間ユニットは，入力を特徴空間に非線形変換するので，多層化という観点からは，カーネル法は事前知識を用いて重みを固定していることになる．しかし，MLPのように線形変換とシグモイド関数に限定せず，一般の非線形変換を用いることができる．ここで，その計算過程において特徴空間での内積しか現れないことを利用し，非線形写像を明示的に定義する代わりに，内積だけを2変数関数として

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{f}'(\boldsymbol{x}_1)\boldsymbol{f}(\boldsymbol{x}_2)$$

のように定義しておくのがカーネル法である．一般に，M 次元空間での内積には M に比例する計算量が必要であるが，内積をカーネル関数 K で定義しておけば，特徴空間の次元が高くても計算量は増加しない．これはカーネルトリックと呼ばれる．また，特徴空間を意識せずにカーネル関数を決定できるというメリットもある．実際，マーサーの定理により，カーネル関数 K が連続で非負定値ならば，K を内積として持つような特徴空間が存在することが証明されている．

以上のことから，明示的に特徴空間を利用しなくても，カーネル法は特徴空間を利用した線形二分機械である．一方，前述のように線形二分機械の平均汎化誤差は入力の次元に比例する．したがって，カーネル法では特徴空間の次元が高いと汎化能力は低下すると予想される．実際，統計力学による平均汎化誤差解析の結果は，この予想を支持している．一方，平均汎化誤差の上限を与えるPAC学習

の枠組みでは，汎化誤差は特徴空間の次元の高さに影響されないことが知られている．これは明らかに矛盾している．以下ではその原因について説明する．

線形二分機械とカーネル法の違いは，入力の分布である．すなわち，線形二分機械では，通常，入力は入力空間全体に分布することが仮定される．一方カーネル法では，特徴ベクトルはそもそも入力ベクトルが写像されたものであるから，入力空間が m 次元ならば特徴ベクトルは特徴空間 F 内の m 次元多様体に局在し，実質的な次元は必ずしも特徴空間の次元と一致しない．では，カーネル法において，特徴空間の実質的な次元はいくつであろうか．これを明らかにするため，入力が 1 次元である多項式カーネル二分機械について考察しよう．すなわち，入力 $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)' \in \boldsymbol{R}^m$ に対し，出力は $y = \mathrm{sgn}\,[\boldsymbol{a}'\boldsymbol{f}(\boldsymbol{x})] \in \{\pm 1\}$ であるとする．ここで $\boldsymbol{f}$ は

$$\boldsymbol{f}(\boldsymbol{x}) = \{c_{d_1,d_2,\ldots,d_m} x_1^{d_1} x_2^{d_2} \cdots x_m^{d_m} | \sum_{i=1}^{m} d_i \leq p\} \in \boldsymbol{R}^M$$

であり，$M = {}_{m+p}\mathrm{C}_m$, $c_{d_1,\ldots,d_m}$ は $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}'\boldsymbol{x} + 1)^p$ で定まる非零定数である．例題の出力は，真の重みベクトル $\boldsymbol{a}_\mathrm{o}$ により $y_n = \mathrm{sgn}\,[\boldsymbol{a}_\mathrm{o}'\boldsymbol{x}_n]$, $n = 1, 2, \ldots, N$ で与えられる．

最も簡単な場合，すなわち $m = 1$ の時，分離多項式は 1 変数で因数分解可能となる．すなわち，分離超平面は

$$\boldsymbol{a}'\boldsymbol{f}(\boldsymbol{x}) = \sum_{i=0}^{p} a_i x^i = c \prod_{i=1}^{p_\mathrm{o}} (x - z_i) \prod_{i=1}^{(p-p_\mathrm{o})/2} (x^2 + b_i x + c_i) = 0$$

と表される．これは，真の分離多項式 $\boldsymbol{a}_\mathrm{o}'\boldsymbol{f}(\boldsymbol{x})$ の零点数は $p_\mathrm{o} (< p)$, 言い換えれば，p 次多項式で分離しようとしているが，真の多項式は p_o 次式であることを意味している．この事実により，p 次多項式の問題は p_o 個の 1 次式の問題に分割される．1 次式の問題では Gibbs アルゴリズムの平均汎化誤差は $2/(3N)$ であることが知られているので，この場合の平均汎化誤差は

$$\varepsilon(N) = \frac{2p_\mathrm{o}}{3N}$$

となり，$\varepsilon(N)$ は p には依存しない [6].

上の議論の一部は $m \geq 2$ の場合にも拡張することができ，分離多項式のイデアルを考えることで，やはりカーネル法による平均汎化誤差は，特徴空間を入力とする線形二分機械の平均汎化誤差よりも小さいことが示される [7].

9 サポートベクトルマシンの幾何学

サポートベクトルマシンは，マージン最大化により定式化される．まず，$\boldsymbol{a}'\boldsymbol{x} + b = 0$ とした場合の定数倍自由度を除去するため，最も近い例題を通る超平面を

$\boldsymbol{a}'\boldsymbol{x}+b=1$ および $\boldsymbol{a}'\boldsymbol{x}+b=-1$ とする (図 7). 例題 $\boldsymbol{x}_1, \boldsymbol{x}_2$ を用いると，マージン最大化は

$$\frac{\boldsymbol{a}'}{2\|\boldsymbol{a}\|}(\boldsymbol{x}_1-\boldsymbol{x}_2)=\frac{1}{\|\boldsymbol{a}\|}\to \max$$

であるので，SVM は

$$\min_{\boldsymbol{a}} \frac{1}{2}\|\boldsymbol{a}\|^2 \quad \text{s.t.} \quad y_n(\boldsymbol{a}\boldsymbol{x}_n+b)\geq 1$$

と定式化される.

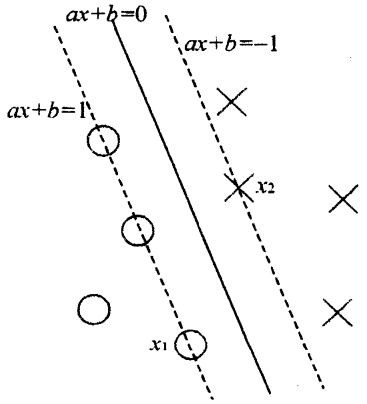


図 7: マージン最大化による SVM の定式化.

一般に，上の 2 次計画問題を直接解く代わりに，その双対問題を用いる．双対問題は，

$$\hat{\boldsymbol{a}}=\sum_n \hat{\alpha}_n y_n \boldsymbol{x}_n$$

$$\hat{\boldsymbol{\alpha}}=\arg\max_{\boldsymbol{\alpha}\geq \boldsymbol{0},\boldsymbol{\alpha}'\boldsymbol{y}=0}\left[\sum_n \alpha_n-\frac{1}{2}\sum_{n,m}\alpha_n\alpha_m y_n y_m \boldsymbol{x}_n\boldsymbol{x}_m\right]$$

と表される．最適解においては Karush-Kuhn-Tucker 条件

$$\hat{\alpha}_n[y_n(\hat{\boldsymbol{a}}'\boldsymbol{x}_n+b)-1]=0$$

が成立し，ごく少数の $\hat{\alpha}_n$ だけが非零となることが多い．$\hat{\alpha}_n\neq 0$ なる例題をサポートベクトル (SV) と呼び，これは境界面に最も近い例題である.

さて，SVM の汎化能力解析について考えよう．単純なパーセプトロンでも未解決問題であることからもわかるように，明示的に平均汎化誤差を求めることは困難である．が，簡単な場合について，パラメータが与える影響などを定量的に示すことは可能である．そのため，多少の仮定を導入する．たとえば，非斉次超平

面は扱いにくいので，分離超平面は斉次であると仮定する．これは，リフトアップと呼ばれる技術で実現できる．すなわち，$\boldsymbol{a}'\boldsymbol{x}+b=0$ が $(\boldsymbol{a}',b)(\boldsymbol{x};1)=0$ と表されることを利用する．また，マージンを 1 に固定する従来の SVM の代わりに，マージンを変数 β とし，$-\beta$ を最小化する関数に導入する ν-SVM [15] を解析する．ν-SVM の主問題，双対問題はそれぞれ

$$\min_{\boldsymbol{a}} \frac{1}{2}\|\boldsymbol{a}\|^2-\beta \qquad \text{s.t.} \quad y_n\boldsymbol{a}'\boldsymbol{x}_n \geq \beta$$

$$\hat{\boldsymbol{a}}=\sum_n \hat{\alpha}_n y_n \boldsymbol{x}_n \qquad \hat{\boldsymbol{\alpha}}=\arg\min_{\boldsymbol{\alpha}\geq \boldsymbol{0}} \frac{1}{2}\|\hat{\boldsymbol{a}}\|^2 \qquad \text{s.t.} \quad \sum_n \alpha_n=1$$

と表される．

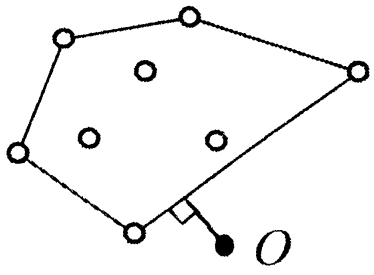


図 8: ν-SVM の幾何学的意味．

ν-SVM の双対問題は，非常にわかりやすい幾何学的意味を持つ．すなわち，$y_n\boldsymbol{x}_n$ の作る凸包で原点に最も近い点が解となる (図 8)．

10 ソフトマージンと縮小凸包

前節で導入した SVM および ν-SVM は，与えられた例題が線形分離不可能な場合には解が存在しない．また，線形分離可能であっても，外れ値などに敏感に反応し，いわゆる過適合を起こすことが知られている．この問題に対処するため，ソフトマージンと呼ばれる手法が提案されている．すなわち，例題が分離超平面よりはみ出すことを許し，超過した大きさ ξ_n を最小化すべき関数に追加する (図 9)．ただし，その重み付けを C とする．すなわち，ν-SVM の主問題は

$$\min_{\boldsymbol{a},\boldsymbol{\xi},\beta} \frac{1}{2}\|\boldsymbol{a}\|^2+C\sum_n \xi_n-\beta \qquad \text{s.t.} \quad y_n\boldsymbol{a}'\boldsymbol{x}_n \geq \beta, \quad \xi_n \geq 0,$$

と表される．

この時，ν-SVM の双対問題は

$$\min_{0\leq\alpha_n\leq C} \frac{1}{2}\|\boldsymbol{a}\|^2 \qquad \text{s.t.} \quad \boldsymbol{a}=\sum_n \alpha_n y_n \boldsymbol{x}_n, \quad \sum_n \alpha_n=1,$$

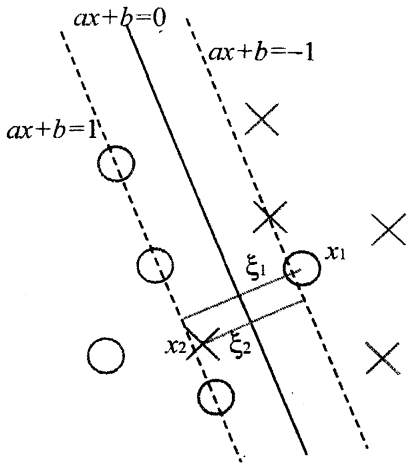


図 9: ソフトマージン.

と表され，これはスラック変数 α_n に上限が与えられた形になっている．すなわち，凸包の一部を除去した集合で，原点に最も近い点が解となる．この集合を縮小凸包と呼ぶ [3]. 自然数 M について，$1/C = M$ の時の縮小凸包は，M 個の例題の重心の集合の凸包と一致する (図 10). すなわち，縮小凸包は例題の分布を反映しており，適切な C では平均化の効果により SN 比が増大し，汎化能力の向上が期待できる一方で，過小な C では例題が中央部に集中し，境界付近の情報が減ることから，汎化能力が低下することが予想される．実際，ν-SVM の幾何学的描像を利用すると，ソフトマージンの効果を定量的に評価することができる [10].

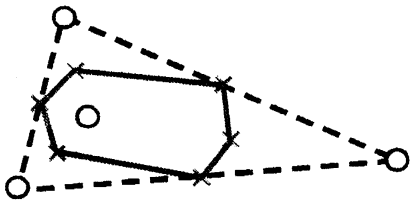


図 10: 縮小凸包の意味 ($C = 1/2$).

一般的な場合については困難なので，ここでは入力は 1 次元半超球面 $\boldsymbol{S}_+$ 上に一様分布すると仮定し，例題数 $N \to \infty$ における平均汎化誤差 $\varepsilon(N)$ を評価する．縮小凸包の性質により，$C = 1/M < 1$ のとき，SVM 解 $\hat{\boldsymbol{a}}$ は，それぞれの端点に最も近い M 個の例題の重心の中点となる (図 11). したがって，重心をそれぞれ θ_{L}, θ_{R} とし，$\theta = (\theta_{\mathrm{L}} - \theta_{\mathrm{R}})/2$ とすれば，$\varepsilon(N) = \langle |\theta|/\pi \rangle$ である．

ある端点に最も近い例題の漸近的分布には，順序統計の考え方を用いることができる．ここで，N が大きいという仮定を利用すると，平均汎化誤差を明示的に

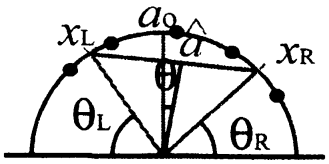


図 11: ν-SVM 解の描像.

導出することができ,

$$\varepsilon_M(N) \approx \frac{1}{NM(M!)^2} \sum_{i,j=1}^{M} \frac{(-1)^{i+j} i^{M+2} j^M {}_M\mathrm{C}_{iM}\mathrm{C}_j}{i+j}$$

である. これでは関数の挙動がわからないので, M がある程度大きいとして中心極限定理と同様の計算を行うと,

$$\varepsilon_M(N) \approx \frac{\sqrt{M}}{\sqrt{3\pi N}}$$

が導かれる.

斉次超平面を持つ ν-SVM の幾何学的描像は, ソフトマージンの他にもカーネル関数の性質と SVM 解の関係の導出にも利用できる [8].

11 まとめ

これまで見てきたように, 脳の情報処理機構を形式的に真似よう, あるいは工学的アプローチで脳の情報処理方式を解明しようという試みは, 徐々にデータからいかに情報を抽出するかという統計科学の課題へと合流してきた. そして統計的視点からの解釈あるいは統計的手法の導入が, アルゴリズムの本質の理解や解析に有用であることが, これまでの研究成果から明らかにされている. 今後, 統計的な手法が有効である分野がますます開拓されていくことであろう.

参考文献

[1] Amari, S.-I.: Theory of Adaptive Pattern Classifiers, *IEEE Trans. on Electronic Computers (EC)*, Vol. 16 (1967), 299–307.

[2] Amari, S.-I. and Murata, N.: Statistical Theory of Learning Curves under Entropic Loss Criterion, *Neural Computation*, Vol. 5 (1993), 140–153.

[3] Bennett, K. P. and Bredensteiner, E. J.: Duality and Geometry in SVM Classifiers, *Proc. Int'l Conf. Machine Learning (ICML)*, (2000), 57–64.

[4] Efron, B.: The Convex Hull of a Random Set of Points, *Biometrika*, Vol. 52 (1965), 331–343.

[5] Fukumizu, K. and Amari, S.: Local Minima and Plateaus in Hierarchical Structures of Multilayer Perceptrons, *Neural Networks*, Vol. 13 (2000), 317–327.

[6] 池田和司：多項式カーネルを持つカーネル法の幾何学と学習曲線, 電子情報通信学会論文誌, Vol. J86-D-II (2003), 918–925.

[7] Ikeda, K.: An Asymptotic Statistical Theory of Polynomial Kernel Methods, *Neural Computation*, Vol. 16 (2004), 1705–1719.

[8] Ikeda, K.: Effects of Kernel Function on Nu Support Vector Machines in Extreme Cases, *IEEE Trans. on Neural Networks (TNN)*, Vol. 17 (2006), 1–9.

[9] Ikeda, K. and Amari, S.-I.: Geometry of Admissible Parameter Region in Neural Learning, *IEICE Trans. Fundamentals*, Vol. E79-A (1996), 938–943.

[10] Ikeda, K. and Aoishi, T.: An Asymptotic Statistical Analysis of Support Vector Machines with Soft Margins, *Neural Networks*, Vol. 18 (2005), 251–259.

[11] McCulloch, W. S. and Pitts, W.: A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bull. Math. Biophys.*, Vol. 5 (1943), 115–133.

[12] Minsky, M. L. and Papert, S. A.: *Perceptrons*, MIT Press, Cambridge, MA, 1969.

[13] Murata, N., Yoshizawa, S. and Amari, S.-I.: Network Information Criterions — Determining the Number of Parameters for an Artifcial Neural Network Model, *IEEE Trans. on Neural Networks (TNN)*, Vol. 5 (1994), 865–872.

[14] Rumelhart, D., McClelland, J. L. and the PDP Research Group, : *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, 1986.

[15] Schölkopf, B., Smola, A. J., Williamson, R. C. and Bartlett, P. L.: New Support Vector Algorithms, *Neural Computation*, Vol. 12 (2000), 1207–1245.

[16] Valiant, L. G.: A Theory of the Learnable, *Communications of ACM*, Vol. 27 (1984), 1134–1142.

[17] Vapnik, V. N.: *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, 1995.

[18] Watanabe, S.: Algebraic Analysis for Nonidentifiable Learning Machines, *Neural Computation*, Vol. 13 (2001), 899–933.