サポートベクターマシンとその応用
An Introduction to Support Vector Machines

山下浩* 田中茂†
Hiroshi Yamashita Shigeru Tanaka
(株) 数理システム‡
Mathematical Systems, Inc.§

概 要

Support Vector Machines (SVMs), the learning approach originally developed by Vapnik and co-workers, have attracted much attention recently because of their excellent performances in various real-world applications such as text categorization, character recognition, image classification, failure discrimination, etc. In this paper we introduce SVMs from the standpoint of mathematical programming applications. Basic properties and various computational aspects of SVMs are described. Our experience on the failure descrimination by using SVMs is given in the final part of this survey.

1 はじめに

近年，データマイニングと言う言葉で括られる一群の手法が広く利用され大きな実用的成果を出している．データマイニングで利用される手法の一つにパターン認識の技術がある．パターン認識の新しい手法としてサポートベクターマシン (support vector machine) が注目を浴びている．注目を浴びる理由はいくつかあるが，主として **(i)** 各種の問題に適用されて優秀な成績をおさめていること，**(ii)** 機械学習手法として興味ある理論的なサポートがあること，**(iii)** 実際問題への適用の際に比較的簡単に実施できること，などがあげられる．また，手法の中心に数理計画が据えられていることは数理計画コミュニティにとっても大いに注目すべきことと思われる．

サポートベクターマシン研究の中心人物は Vladimir N. Vapnik である．Vapnik と彼の共同研究者は 1960 年代から最適超平面による識別法の提案とその汎化能力に関する解析を行っていた．1960 年代には機械学習の分野において，Rosenblatt による単純パーセプトロンの提案や，カーネル法の研究などが同様に行われている．ニューロンの学習モデルに基づいたパーセプトロンは，1980 年代に多層パーセプトロンとその学習アルゴリズムのバックプロパゲーションの導入によって多くの分野で応用されたことは良く知られている．しかし，実際の応用では過剰学習の問題，収束の遅さ，局所最適解への収束，モデルの選択の任意性などのいくつかの問題も指摘されている．

Vapnik 達は 1990 年代になって最適超平面法とカーネル法を組み合わせて，SVM を再度提案しその方法を実用問題に対して適用し，高度にチューニングされたニューラルネットなどに比肩し得

*hy@msi.co.jp
†tanaka@msi.co.jp
‡〒 160-0022 東京都新宿区新宿 2-4-3 フォーシーズンビル 10 階
§10F Four Seasons Bldg. 2-4-3, Shinjuku, Shinjuku-ku, Tokyo, Japan 160-0022

る性能を示し，広く注目を集める近年の状況を招いた．以下で，その基本的性質を順を追って解説したい．また，若干の応用例に関する我々の経験についても最後に紹介したい．

本稿では，主として数理計画の立場からサポートベクターマシンの手法について紹介することを目的としている．その意味で，本サーベイはサポートベクターマシンとその周辺の技法のごく一部の紹介であり，より深くこの分野を研究するためには最後にあげた参考文献を参照されたい[1]．とくに，最近の教科書（[16]，[15]，[1]）や論文集（[13]，[14]）が参考になると思われる．

2 学習

学習（厳密には教師付き学習—supervised learning）を数理的に扱うために，以下のような状況を設定する．学習/トレーニングの対象となるデータ群が与えられたときに，それらの情報から何らかの知見を得て，未知のデータに対する判断を行うことを目的とする．簡単のために，二つのクラスに分けられたデータ群の学習から，未知のデータをどちらかのクラスに分類する作業を考える.（分類クラス数の一般化も可能である.）

トレーニングデータの集合を

$$S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_\ell, y_\ell)\}, \quad \mathbf{x}_i \in \mathbb{R}^N, y_i \in \{-1, 1\}$$

と表す．すなわち，それぞれのデータは N 個の成分（入力）とクラス分けのための指標 $\{-1, 1\}$（出力）から成っている．このデータはある（未知の）分布 $P(\mathbf{x}, y)$ によって生成されていると仮定する．上で述べたように，学習とはこれらのデータから識別関数 $f : \mathbb{R}^N \to \{-1, 1\}$ を選びトレーニングデータに入っていないデータ $(\mathbf{x}, y)$（通常，検証用データと言う）を正しく $(f(\mathbf{x}) = y)$ 推定することを目的とする．与えられたトレーニングデータに対して正しい答えを出す $(f(\mathbf{x}_i) = y_i, i = 1, ..., \ell)$ 関数は無数に存在するし，それらが検証用データに対して異なった答えを出すこともあり得る．したがって，トレーニングデータを正しく識別するという基準だけでは未知のデータに対して正しい答えを与えるという本来の目的（汎化能力—generalization）に沿った関数を選ぶことは出来ない．何らかの基準によって候補となる関数を制限しなくてはならない．

候補とする識別関数の集合を $\mathcal{H}$ とする．期待リスク (expected risk)：

$$R(f) = \int \frac{1}{2} |f(\mathbf{x}) - y| \, P(\mathbf{x}, y) d\mathbf{x} dy$$

を最小にする関数 $f \in \mathcal{H}$ を求めることが究極の目的である．しかし，分布 $P(\mathbf{x}, y)$ は未知なので期待リスク $R(f)$ を計算することは不可能である．そこで，我々は経験的リスク (empirical risk)：

$$R_{emp}(f) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{1}{2} |f(\mathbf{x}_i) - y_i|$$

を扱わざるを得ない．上で述べたように，単純な $R_{emp}(f)$ の最小化（経験的リスク最小化の原理-Empirical Risk Minimization Principle）では自由度があり過ぎるので何らかの基準を必要とする．Vapnik 達による統計的学習理論によると，トレーニングデータの数を増やしていったときに $R_{emp}(f)$ の最小化が $R(f)$ の最小化と矛盾しないためには $R_{emp}(f)$ が $R(f)$ に一様収束することが必要十分条件となる．そして，この条件は集合 $\mathcal{H}$ の VC(Vapnik and Chervonekis)– 次元 h の有界性と同値となる．集合 $\mathcal{H}$ の VC-次元とは，この集合の複雑性/多様性（capacity）を表すもので，「集合に属する関数によってあらゆる可能な分離が出来る最大の点の数」のことである．

[1] `http://www.kernel-machines.org/` で最新の情報や文献にアクセスすることが出来る.

VC–次元によって $R(f)$ と $R_{emp}(f)$ の誤差の評価を行うことができる ([17], [16], [15]). たとえば, $h < \ell$ ならば, 確率 $1-\delta$ で

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h(\log \frac{2\ell}{h} + 1) - \log \frac{\delta}{4}}{\ell}}, \quad \forall f \in \mathcal{H} \tag{1}$$

が成立する[2]. したがって, 期待リスクを小さくするためには上の不等式の右辺の 2 つの項を小さくすることが考えられる. すなわち, $R_{emp}(f)$ と h/ℓ を小さくすれば良い. 一般に, $R_{emp}(f)$ の値は h (関数の自由度) が大きい程小さくなるので, 上記右辺の和を最小にするような VC–次元の最適値が存在するであろう.

Vapnik による構造的リスク最小化 (Structural Risk Minimization) の原理は VC–次元 h_n が単調増加するような関数集合の列

$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset ... \subset \mathcal{H}_n \subset ...$$

を考えて, 上記不等式の右辺, あるいは簡単化して $R_{emp}(f) + \sqrt{h_n/\ell}$ を最小化するような $\mathcal{H}_n$ と $f \in \mathcal{H}_n$ を求めようとするものである. 厳密にこの操作を実行するのは困難であるが, サポートベクターマシンはある意味でこの原理をなぞっているものと解釈できる.

3 超平面による識別

トレーニングデータ集合 S が $\mathbb{R}^N$ の超平面によって $y_i = 1$ のグループと $y_i = -1$ のグループに分離される場合を線形分離可能 (linearly separable) と言う. 候補となる超平面を

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0 \tag{2}$$

と表す. ここで, $\mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}$ で, $(\mathbf{w} \cdot \mathbf{x})$ は $\mathbf{w}$ と $\mathbf{x}$ の内積を表す. そして, 識別関数を

$$f(x) = \mathrm{sgn}\left((\mathbf{w} \cdot \mathbf{x}) + b\right) \tag{3}$$

とする. ($\mathrm{sgn}(0) = 1$ と約束する.) 一般にトレーニングデータを分離する超平面は一意には決まらないことに注意する.

超平面 $(\mathbf{w}, b)$ に対するサンプル $(\mathbf{x}_i, y_i)$ のマージンは

$$\gamma_i = \frac{y_i\left((\mathbf{w} \cdot \mathbf{x}_i) + b\right)}{\|\mathbf{w}\|}$$

と定義される. ($\|\cdot\|$ は 2 ノルムを表す.) $\gamma_i > 0$ ならば $(\mathbf{x}_i, y_i)$ は正しく識別されていることになる. 点 $\mathbf{x} \in \mathbb{R}^N$ を超平面 (2) へ射影した点を $\widehat{\mathbf{x}} \in \mathbb{R}^N$ とすると,

$$\|\mathbf{w}\|\,\|\mathbf{x} - \widehat{\mathbf{x}}\| = |(\mathbf{w} \cdot (\mathbf{x} - \widehat{\mathbf{x}}))| = |(\mathbf{w} \cdot \mathbf{x} + b) - (\mathbf{w} \cdot \widehat{\mathbf{x}} + b)| = |(\mathbf{w} \cdot \mathbf{x} + b)|$$

となるから, $\gamma_i > 0$ ならば γ_i は点 $\mathbf{x}_i$ から超平面 $(\mathbf{w}, b)$ への距離を表す.

上記マージン $\gamma_i, i = 1, ...\ell$ の最小値をサンプルデータ集合 S に対する超平面 $(\mathbf{w}, b)$ のマージンと言う. そして, すべての超平面のマージンの最大値をこのデータ集合 S のマージンと呼ぶ. 線形分離可能なデータ集合に対してはマージンは正となる.

[2] このような評価を PAC(probably approximately correct) という.

3.1 線形分離可能な場合

以下では，線形分離可能なデータ集合を考える．この場合，超平面 $(\mathbf{w}, b)$ が存在して，

$$y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)>0,\ \ i=1,...\ell$$

となるから，超平面を正規化して

$$\min_i\left\{y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)\right\}=1$$

とすることが出来る．この性質をみたす形式を正準形 (canonical form) と言う[3]．正準形式の超平面のマージンは $1/\|\mathbf{w}\|$ となる．マージンが最大になるような超平面を最適な超平面 (optimal hyperplane) と考えると，それは

$$\begin{array}{ll} \text{最小化} & \frac{1}{2}\|\mathbf{w}\|^2, \qquad\qquad \mathbf{w}\in\mathbb{R}^N, b\in\mathbb{R} \\ \text{条　件} & y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)\geq 1, \quad i=1,...\ell \end{array} \tag{4}$$

の解によって与えられる．

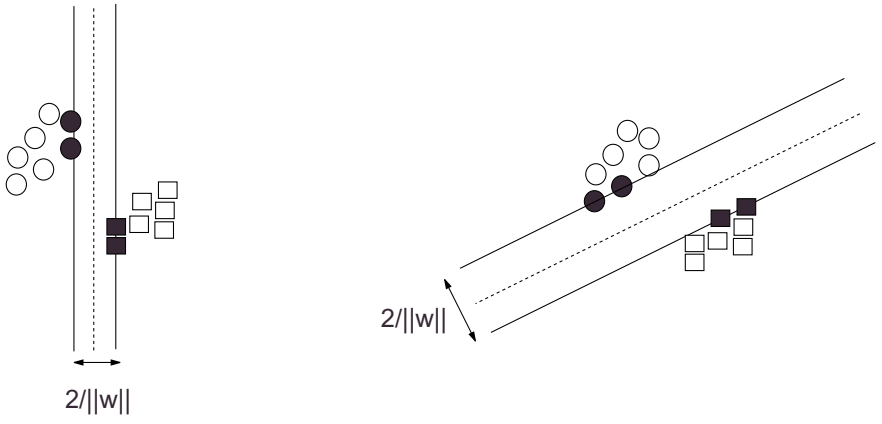

図 1: 分離超平面

この問題 (4) は凸 2 次計画問題であり，問題の構造も特に複雑ではないが，以下で非線形の（曲面による）識別に拡張するために双対問題を考えると都合が良い．ラグランジュ関数を $L(\mathbf{w}, b, \boldsymbol{\alpha})$ とする．ここで，$\boldsymbol{\alpha}\in\mathbb{R}^\ell$ は双対変数である．Wolfe の双対問題は

$$\begin{array}{ll} \text{最大化} & L(\mathbf{w},b,\boldsymbol{\alpha}), \qquad \mathbf{w}\in\mathbb{R}^N, b\in\mathbb{R}, \boldsymbol{\alpha}\in\mathbb{R}^\ell \\ \text{条　件} & \nabla_{\mathbf{w}}L(\mathbf{w},b,\boldsymbol{\alpha})=0, \nabla_b L(\mathbf{w},b,\boldsymbol{\alpha})=0, \boldsymbol{\alpha}\geq\mathbf{0} \end{array}$$

となる．具体的には

$$L(\mathbf{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{\ell}\alpha_i\left(y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)-1\right)$$

であるから，最適性の必要十分条件である Karush-Kuhn-Tucker(KKT) 条件は

$$\nabla_{\mathbf{w}}L(\mathbf{w},b,\boldsymbol{\alpha}) \quad = \quad \mathbf{w}-\sum_{i=1}^{\ell}\alpha_i y_i\mathbf{x}_i=0 \tag{5}$$

[3]正準超平面の集合の VC–次元は $N+1$ であることが知られている ([16]).

$$\nabla_b L(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i y_i = 0, \tag{6}$$

$$\alpha_i \left(y_i \left((\mathbf{w} \cdot \mathbf{x}_i) + b \right) - 1 \right) = 0, \ \ \alpha_i \geq 0, \ y_i \left((\mathbf{w} \cdot \mathbf{x}_i) + b \right) - 1 \geq 0, \ \ i = 1, ...\ell \tag{7}$$

これらの条件を考慮すると，双対問題は

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^{\ell} \alpha_i, \quad \boldsymbol{\alpha} \in \mathbb{R}^{\ell} \\ \text{条　件} & \sum_{i=1}^{\ell} \alpha_i y_i = 0, \boldsymbol{\alpha} \geq \mathbf{0} \end{array} \tag{8}$$

となる．双対問題のヘッセ行列を $-\mathbf{D} \in R^{\ell \times \ell}$, $D_{ij} = y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ と置くと行列 $\mathbf{D}$ は対称非負定値行列となり，上記問題は

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2} (\boldsymbol{\alpha} \cdot \mathbf{D}\boldsymbol{\alpha}) + (\mathbf{e} \cdot \boldsymbol{\alpha}), \quad \boldsymbol{\alpha} \in \mathbb{R}^{\ell} \\ \text{条　件} & (\mathbf{y} \cdot \boldsymbol{\alpha}) = 0, \boldsymbol{\alpha} \geq \mathbf{0} \end{array}$$

と書ける．ここで，$\mathbf{e} = (1, ..., 1)^t \in \mathbb{R}^{\ell}$ である．

最適解を $(\mathbf{w}^*, b^*, \boldsymbol{\alpha}^*)$ とすると，双対変数 $\alpha_i^*, i = 1, ...\ell$ の中で非零のものは不等式制約条件が有効となっているもの，すなわち

$$y_i \left((\mathbf{w}^* \cdot \mathbf{x}_i) + b \right) = 1 \tag{9}$$

となるものなので，数が少ない場合が多い．(5) より，最適解を与える $\mathbf{w}^*$ は非零の α_i^* に対応するサンプルデータ $\mathbf{x}_i$ の 1 次結合によって表される．そのようなデータをサポートベクターと呼ぶ．サポートベクターの集合を $S_V \subset S$ と表わすと，

$$\mathbf{w}^* = \sum_{i \in S_V} \alpha_i^* y_i \mathbf{x}_i$$

である．双対変数のほんの一部のみが非零となる（有効制約が少数である）ことはスパース性 (sparseness) と呼ばれている．(9) より，b^* も任意のサポートベクター $\mathbf{x}_i$ によって

$$b^* = y_i - (\mathbf{w}^* \cdot \mathbf{x}_i) = y_i - \sum_{j \in S_V} \alpha_j^* y_j (\mathbf{x}_j \cdot \mathbf{x}_i)$$

と表される．識別関数は

$$f(x) = \mathrm{sgn} \left(\sum_{i \in S_V} \alpha_i^* y_i \left(\mathbf{x} \cdot \mathbf{x}_i \right) + b^* \right)$$

となる．また，b^* の表式より

$$\|\mathbf{w}^*\|^2 = \sum_{i, \in S_V} \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{w}^*) = \sum_{i \in S_V} \alpha_i^* (1 - y_i b^*) = \sum_{i \in S_V} \alpha_i^*$$

となるのでデータ集合 S のマージン γ は

$$\gamma = \frac{1}{\|\mathbf{w}^*\|} = \left(\sum_{i \in S_V} \alpha_i^* \right)^{-1/2}$$

で与えられる．

3.2 線形分離が不可能な場合

次に，線形分離が不可能 (linearly non-separable) な場合，すなわちデータ集合 S のマージンが正にならない場合を考える．この場合は元の問題の条件：

$$\text{条　件}\quad y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)\geq 1,\quad i=1,...\ell$$

にスラック変数を導入して

$$\text{条　件}\quad y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)\geq 1-\xi_i,\xi_i\geq 0\quad i=1,...\ell$$

と変換して，制約条件がすべてみたされない場合に対応する．このような形式の制約条件を扱うものをソフトマージンによる最適化と言う．スラック変数の値はなるべく小さくしたいので目的関数にペナルティを課すのは最適化の常套手段であろう．ペナルティ項の形の代表的なものを二つあげておく．

3.2.1 1 ノルムによるソフトマージン最適化

解くべき問題を

$$\begin{array}{ll}\text{最小化} & \frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{\ell}\xi_i, \qquad\qquad \mathbf{w}\in\mathbb{R}^N, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^\ell\\ \text{条　件} & y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)\geq 1-\xi_i,\xi_i\geq 0,\quad i=1,...\ell\end{array}$$

とする．ここで，$C>0$ はペナルティパラメータである[4]．ラグランジュ関数を

$$L(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{\ell}\xi_i-\sum_{i=1}^{\ell}\alpha_i\left(y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)-1+\xi_i\right)-\sum_{i=1}^{\ell}\beta_i\xi_i$$

とする．ここで，$\boldsymbol{\alpha}\in\mathbb{R}^\ell$ と $\boldsymbol{\beta}\in\mathbb{R}^\ell$ は双対変数である．線形分離可能な場合と同様に，KKT 条件は

$$\nabla_{\mathbf{w}}L(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathbf{w}-\sum_{i=1}^{\ell}\alpha_i y_i\mathbf{x}_i=\mathbf{0} \tag{10}$$

$$\nabla_b L(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{i=1}^{\ell}\alpha_i y_i=0 \tag{11}$$

$$\nabla_{\boldsymbol{\xi}}L(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = C\mathbf{e}-\boldsymbol{\alpha}-\boldsymbol{\beta}=\mathbf{0} \tag{12}$$

$$\alpha_i\left(y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)-1+\xi_i\right) = 0,\ \alpha_i\geq 0,\ y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)-1+\xi_i\geq 0,\quad i=1,...\ell \tag{13}$$

$$\beta_i\xi_i = 0,\ \beta_i\geq 0,\ \xi_i\geq 0,\ i=1,...\ell \tag{14}$$

となる．(12) から得られる $\boldsymbol{\beta}=C\mathbf{e}-\boldsymbol{\alpha}$ を (14) に代入すると

$$(C-\alpha_i)\xi_i=0,\ \ \alpha_i\leq C,\ \xi_i\geq 0,\ i=1,...\ell \tag{15}$$

を得る．結局，双対問題から $\mathbf{w}$,b,$\boldsymbol{\xi}$,$\boldsymbol{\beta}$ が消去できて

$$\begin{array}{ll}\text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i\cdot\mathbf{x}_j)+\sum_{i=1}^{\ell}\alpha_i,\quad \boldsymbol{\alpha}\in\mathbb{R}^\ell\\ \text{条　件} & \sum_{i=1}^{\ell}\alpha_i y_i=0, \mathbf{0}\leq\boldsymbol{\alpha}\leq C\mathbf{e}\end{array} \tag{16}$$

を得る．分離可能の場合と異なるのは，変数に上限が付加されることである．

[4]実際の計算では，$y_i=1$ のサンプルには $C_+>0$，$y_i=-1$ のサンプルには $C_-<0$，と異なったペナルティパラメータを与えることが多い．このことは，以下に現れる各種の定式化でも同様である．

- 双対変数 $\alpha_i^* > 0$ に対応するのは，不等式制約条件が有効となっているもの，すなわち

$$y_i\left((\mathbf{w}^* \cdot \mathbf{x}_i) + b^*\right) = 1 - \xi_i^*$$

であり，さらに $\alpha_i^* < C$ ならば $\xi_i^* = 0$ となる．したがって，$0 < \alpha_i^* < C$ となるデータが正しく識別されたサポートベクターとなる：

$$0 < \alpha_i^* < C \Longrightarrow y_i\left((\mathbf{w}^* \cdot \mathbf{x}_i) + b^*\right) = 1$$

- $\alpha_i^* = 0$ の場合は，(15) より $\xi_i = 0$ となるので，

$$\alpha_i^* = 0 \Longrightarrow y_i\left((\mathbf{w}^* \cdot \mathbf{x}_i) + b^*\right) \geq 1$$

- $\alpha_i^* = C$ の場合は，$y_i\left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) - 1 + \xi_i = 0$ かつ $\xi_i \geq 0$ となるので，

$$\alpha_i^* = C \Longrightarrow y_i\left((\mathbf{w}^* \cdot \mathbf{x}_i) + b^*\right) \leq 1$$

上記より，正しく識別されたサポートベクター $\mathbf{x}_i$ が存在するとき b^* は

$$b^* = y_i - (\mathbf{w}^* \cdot \mathbf{x}_i) = y_i - \sum_{j \in S_V} \alpha_j^* y_j (\mathbf{x}_j \cdot \mathbf{x}_i)$$

と表される．

3.2.2 2 ノルムによるソフトマージン最適化

2 ノルムのペナルティを利用する場合には，問題は

$$\begin{array}{ll} \text{最小化} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{\ell}\xi_i^2, \qquad \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^\ell \\ \text{条　件} & y_i\left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) \geq 1 - \xi_i, \quad i = 1, ...\ell \end{array}$$

となる．ここで，ξ_i に対する非負条件は不必要なので除去されている．何故なら，$\xi_i < 0$ の点では $y_i\left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) \geq 1$ の制約条件はみたされていて，$\mathbf{w}, b$ の値を固定したまま $\xi_i = 0$ とした方が目的関数の値はより小さくなる．したがって，$\xi_i < 0$ が最適解として得られることは無いからである．ラグランジュ関数は

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{\ell}\xi_i^2 - \sum_{i=1}^{\ell}\alpha_i\left(y_i\left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) - 1 + \xi_i\right)$$

となる．KKT 条件は

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{i=1}^{\ell}\alpha_i y_i \mathbf{x}_i = \mathbf{0} \tag{17}$$

$$\nabla_b L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \sum_{i=1}^{\ell}\alpha_i y_i = 0 \tag{18}$$

$$\nabla_\xi L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = C\boldsymbol{\xi} - \boldsymbol{\alpha} = \mathbf{0} \tag{19}$$

$$\alpha_i\left(y_i\left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) - 1 + \xi_i\right) = 0, \ \ \alpha_i \geq 0, \ y_i\left((\mathbf{w} \cdot \mathbf{x}_i) + b\right) - 1 + \xi_i \geq 0, \ \ i = 1, ...\ell \tag{20}$$

となる．上式より，双対問題から $\mathbf{w}$,b,$\boldsymbol{\xi}$ が消去できて

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j\left((\mathbf{x}_i\cdot\mathbf{x}_j)+\frac{1}{C}\delta_{ij}\right)+\sum_{i=1}^{\ell}\alpha_i, \quad \boldsymbol{\alpha}\in\mathbb{R}^{\ell} \\ \text{条　件} & \sum_{i=1}^{\ell}\alpha_i y_i=0, \boldsymbol{\alpha}\geq\mathbf{0} \end{array}$$

を得る．ここで，δ_{ij} は Kronecker のデルタ記号である．1 ノルムの場合と異なるのは，変数 $\boldsymbol{\alpha}$ に関する上限制約が存在しないことと，ヘッセ行列の対角項に $1/C$ という数が引かれていることである．したがって，この場合，ヘッセ行列は負定値となり，数値計算上はより好条件となる．

最適解では，

- $\alpha_i^*=0$ ならば $\xi_i^*=0$ となり，

$$\alpha_i^*=0\Longrightarrow y_i\left((\mathbf{w}^*\cdot\mathbf{x}_i)+b^*\right)\geq 1$$

- $\alpha_i^*>0$ ならば $\xi_i^*=\alpha_i^*/C>0$ なので，

$$\alpha_i^*>0\Longrightarrow y_i\left((\mathbf{w}^*\cdot\mathbf{x}_i)+b^*\right)<1$$

b^* は任意のサポートベクター $\mathbf{x}_i$ によって

$$b^*=y_i(1-\alpha_i^*/C)-(\mathbf{w}^*\cdot\mathbf{x}_i)=y_i(1-\alpha_i^*/C)-\sum_{j\in S_V}\alpha_j^* y_j(\mathbf{x}_j\cdot\mathbf{x}_i)$$

と計算される．

3.2.3　正準超平面の性質

分離超平面を構成した正準超平面は（2）で定義されているが,「$\mathbb{R}^N$ の正準超平面全体の集合の VC– 次元は $N+1$ となる」ことが知られている．これらの候補の中から構造的リスク最小化原理の考え方に沿ったストラテジーを実施するためには，経験的リスクと VC–次元の両方の最小化を視野に入れる必要がある．そのために，正準超平面の集合から VC–次元の異なる部分集合を生成する必要がある．Vapnik はその観点から次のような興味ある定理を証明している．

定理（正準超平面の VC–次元）　データ集合 S の点 $x_1,...,x_\ell$ が $\mathbb{R}^N$ 内の半径 R の球に含まれるとき，

$$\|\mathbf{w}\|\leq A$$

をみたす正準超平面の部分集合の VC–次元は

$$h\leq\min\left\{\left[R^2A^2\right],N\right\}+1$$

となる．ここで，$\left[R^2A^2\right]$ は R^2A^2 の整数部分を意味する．　□

この定理から分かるように，$\|\mathbf{w}\|$ を小さくすること（マージン最大化）と正準超平面の部分集合の VC– 次元を小さくすることは同値である．ソフトマージン最適化のペナルティ項は経験的リスクを小さくする項であるから，ソフトマージン最適化問題は構造的リスク最小化原理の方針に従った解法であることが分かる．

また，線形分離可能なデータ集合に対して以下のような定理が証明されている．

定理（誤り率の期待値）　線形分離可能な ℓ 個のトレーニングデータ集合に対して最適分離超平面を生成するとき，この識別関数の誤り率の期待値 $E(P_{error})$ は次のような上界を持つ：

$$E(P_{error}) \leq \frac{E(\min\left\{m, D^2/\gamma^2, N\right\})}{\ell}$$

ここで，m はサポートベクターの数，$D = \max_i \|\mathbf{x}_i\|$，$\gamma$ はマージンを，それぞれ表す確率変数である．□

この定理から，誤り率を減らすことはマージンを大きくすることと，サポートベクターの数を減らすことに関連していることが分かる．上の評価式には入力空間の次元 N が現れているが，$\min\{\cdot\}$ の中にあるので，通常は N が他の項と比べて大きいとき評価に影響しないことに注意する．機械学習において，入力空間の次元（あるいは次節で述べるような特徴空間への変換後の次元）が大きくなると汎化能力が低下する問題（次元の呪い–curse of dimensionality）が知られているが，SVM ではそれが回避されていることになる．この種の評価式の様々な形のものが [1] に解説されている．

4　曲面による識別

すべての場合に平面による識別が適切とは限らない．そこで，より複雑な識別に対応するために曲面による分離を考える．このような場合に，高次元の空間への非線形変換とその空間でのカーネルトリックと言われる方法が知られている．まず，入力データをより高次元な空間に（非線形）射像する．すなわち，

$$\mathbf{x} \mapsto \boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), ...)$$

と対応付けられる入力データ空間 $X \subset \mathbb{R}^N$ から特徴空間（feature space）$F = \{\boldsymbol{\phi}(\mathbf{x}) \,|\mathbf{x} \in \mathbf{X}\}$ への射像を考える．そして，特徴空間において線形分離を考える．その際に，双対問題で表された最

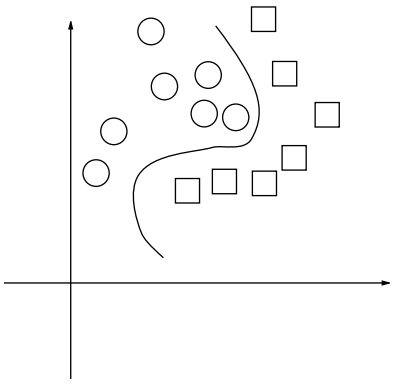

図 2: 入力空間での分離

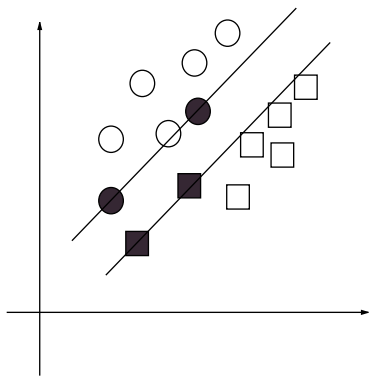

図 3: 特徴空間での分離

適化問題において，特徴空間に対応した量はすべて内積の形式でのみ現れることに注意する．すなわち，$(\boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{y})), \mathbf{x}, \mathbf{y} \in X$ の形式である．高次元空間において内積を文字通り計算するのは現実的ではないので，対応する項をカーネル関数で置き換える：

$$K(\mathbf{x}, \mathbf{y}) = (\boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{y}))$$

このとき，識別関数は

$$f(x) = \mathrm{sgn}\left(\sum_{i \in S_V} \alpha_i^* y_i K(\mathbf{x}, \mathbf{x}_i) + b^*\right) \tag{21}$$

となる．

1 ノルムソフトマージン最適化問題は

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^{\ell} \alpha_i, \quad \boldsymbol{\alpha} \in \mathbb{R}^{\ell} \\ \text{条　件} & \sum_{i=1}^{\ell} \alpha_i y_i = 0,\ 0 \leq \boldsymbol{\alpha} \leq C\mathbf{e} \end{array}$$

となり，

$$b^* = y_i - \sum_{j \in S_V} \alpha_j^* y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

また，2 ノルムソフトマージン最適化問題は

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \left(K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{C}\delta_{ij}\right) + \sum_{i=1}^{\ell} \alpha_i, \quad \boldsymbol{\alpha} \in \mathbb{R}^{\ell} \\ \text{条　件} & \sum_{i=1}^{\ell} \alpha_i y_i = 0,\ \boldsymbol{\alpha} \geq \mathbf{0} \end{array}$$

となり，

$$b^* = y_i(1 - \alpha_i^*/C) - \sum_{j \in S_V} \alpha_j^* y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

任意に与えた関数 $K(\mathbf{x}, \mathbf{y})$ をカーネル関数として使えるわけではない．それは，内積の形で表せる必要があるが，その保証を与えるのが，関数解析で古くから知られている Mercer の定理である．

定理　(Mercer)　X は $\mathbb{R}^N$ の有界閉集合で，関数 $K : X \times X \to \mathbb{R}$ は連続かつ対称（$K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$）とする．このとき，関数 K が，一様収束する級数：

$$K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} a_j \phi_j(\mathbf{x}) \phi_j(\mathbf{y}), \quad a_j > 0$$

によって展開可能となる必要十分条件は

$$\int_{X \times X} K(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0, \quad \forall f \in L_2(X) \tag{22}$$

である．　□

この定理の正定符号カーネルの条件 (22) は

$$\sum_{i,j=1}^{n} K(\mathbf{x}_i, \mathbf{x}_j) v_i v_j \geq 0, \ \ \forall \mathbf{x}_i, \mathbf{x}_j \in X, \forall v_i, v_j \in \mathbb{R}$$

と表すことができる．すなわち，行列 $\{K(\mathbf{x}_i, \mathbf{x}_j), i, j = 1, ..., n\}$ が非負定値という条件である．

Mercer の条件をみたす具体的なカーネル関数の形としてよく使用されているものをあげておく．

- radial basis function(RBF) カーネル　　$K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|/(2\sigma^2))$
- d 次の多項式カーネル　　$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^d$
- シグモイドカーネル　　$K(\mathbf{x}, \mathbf{y}) = \tanh(\kappa(\mathbf{x} \cdot \mathbf{y}) - \boldsymbol{\theta})$

下図に（S-PLUS による）線形判別，ニューラルネット，SVM（RBF カーネル）を用いた識別例を示す．

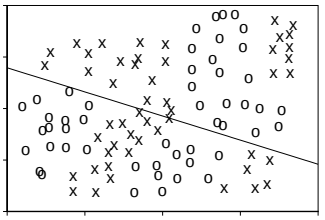

図 4: 線形判別

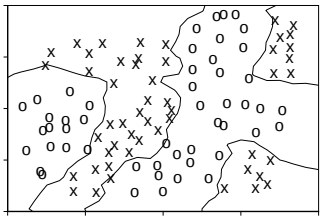

図 5: ニューラルネット

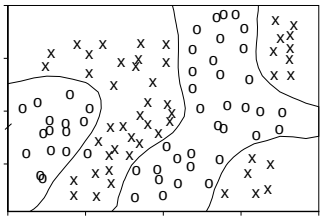

図 6: SVM

5　サポートベクターマシンのアルゴリズム

本節では，最も広く用いられている 1 ノルムソフトマージン最適化問題：

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)+\sum_{i=1}^{\ell}\alpha_i, \quad \boldsymbol{\alpha}\in\mathbb{R}^{\ell} \\ \text{条　件} & \sum_{i=1}^{\ell}\alpha_i y_i = 0, \mathbf{0}\le\boldsymbol{\alpha}\le C\mathbf{e} \end{array} \tag{23}$$

を解くためのアルゴリズムについて概観する．2 ノルム問題についても以下の議論はほとんど同様に適用される．1 ノルム問題と異なる点は，ヘッセ行列が負定値になることと変数に上限制約が存在しないことである．

5.1　問題の性質

問題 (23) は簡単な構造の 2 次計画問題であるが，以下のような特徴を持っている．

1. ヘッセ行列は非正定値である．したがって，アルゴリズムはヘッセ行列が特異になる場合を考慮する必要がある．
2. 制約条件は，1 本の等式制約と変数に対する上下限制約のみである．
3. 問題の規模は学習サンプル数が多いとき大規模になる．変数の数は学習サンプルの数に一致する．ヘッセ行列は密行列となるから，大規模な学習に対しては特別な配慮が必要になる．カーネル $K(\mathbf{x}_i,\mathbf{x}_j)$ をどのように保持するか（あるいは明示的には保持しないか），計算にどのように利用するか考慮の必要な点である．
4. 最適解では多くの変数が下限値 (0) にある．アルゴリズムでこの性質をどのように利用するかが重要な点となる．

以下では，このような点に留意していくつかのアルゴリズムについて考えてみる．

5.2　汎用 2 次計画問題ソルバー

学習サンプル数がそれほど多くないときは，SVM の解法として一般の QP ソルバーを使うことは可能である．ただし，ヘッセ行列が特異行列になることを考慮した解法である必要がある．汎用アルゴリズムとして，内点法と有効制約法があるが，内点法はサポートベクターの数が少ない（主問題での有効制約の数が少ない/双対問題では多くの変数が下限値にある）ことを利用できな

い．反復の初期値を内点にとる（下限値に取れない）ことと，有効な制約と有効でない制約を明確に判定することが不得意であることが，問題である．

このような点を考慮すると汎用アルゴリズムとしては有効制約法がより適していると言える．その場合，初期値を 0 に取ることが最適解の良い近似になっているはずである．双対問題で下限値から離れる変数の数が少ないことから，有効制約の係数行列の零空間の基底を保持して反復する方法 (null space method) が適していると思われる．

しかし，零空間基底を利用した有効制約法でも，ヘッセ行列が密であることから大規模学習サンプルへの対応は難しい．下限値に止まっている変数と対応するカーネルの成分 $K(\mathbf{x}_i, \mathbf{x}_j)$ を計算に明示的に持ち込まないようにする工夫が必要になる．そのような立場から，既にいくつかのヒューリスティックなアルゴリズムが実用化されている．

5.3 chunking と decomposition

Vapnik による chunking と呼ばれる戦略が最も簡単なものであろう．それは，以下のような手順で実施される．全学習データから一部（chunk）を取り出して，QP ソルバーによって対応する問題（working set とも呼ぶ）を解く．得られたサポートベクターを使って残りのデータに対してテストを行う．誤った識別をしたデータ（$y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)-1\geq 0$ をみたしていないもの）の中から，外れ度の大きいものから一定個取り出して前回のサポートベクターと合わせて新しい問題とする．前回の最適解を利用した初期値から出発して QP を解く．以上を繰り返して，最後に正しいサポートベクターを同定して終了するのが chunking と呼ばれる方法である．この方法では，学習データによっては，比較的大きな（密）QP 問題を解くことが必要になる可能性があるので，実用性には制限があると思われる．

decomposition method は QP のサイズが大きくなるのを防ぐために，working set の大きさをあらかじめ制限してしまう方法である．サポートベクターの候補となる変数をすべて QP の変数として扱わずにその一部のみを動かすことによって，問題のサイズを小規模に止めて反復を繰り返す．このクラスの方法（コード）の代表的なものに SVM^{light} と SMO がある．

5.3.1 SVM^{light}

まず，問題の decomposition を以下のように定義する．変数を自由に動ける集合 B と，値の固定された集合 N に分ける．B には q 個の変数があり，残りの $\ell-q$ 個は N に入っている．すると，解くべき問題は

$$\begin{array}{ll} \text{最大化} & -\frac{1}{2}\sum_{i,j\in B}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)+\sum_{i\in B}\alpha_j\left(1-\sum_{i\in N}\alpha_i y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)\right) \\ & -\frac{1}{2}\sum_{i,j\in N}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)+\sum_{i\in N}\alpha_j \\ \text{条　件} & \sum_{i\in B}\alpha_i y_i+\sum_{i\in N}\alpha_i y_i=0,\ \mathbf{0}\leq\boldsymbol{\alpha}\leq C\mathbf{e} \end{array} \tag{24}$$

となる．$\alpha_j, j\in N$ は定数である．

SVM^{light} では各反復において集合 B を目的関数がなるべく大きく増大するように選ぶ．そのた

めに，まず上の QP を線形近似した以下のような問題を解く．

$$\begin{array}{ll} \text{最大化} & -\sum_{i,j=1}^{\ell} \alpha_i y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) d_j + \sum_{i=1}^{\ell} d_i, \quad \mathbf{d} \in \mathbb{R}^{\ell} \\ \text{条　件} & \sum_{i=1}^{\ell} d_i y_i = 0, \\ & d_j \geq 0, \ (\alpha_j = 0) \\ & d_j \leq 0, \ (\alpha_j = C) \\ & -\mathbf{e} \leq \mathbf{d} \leq \mathbf{e} \end{array}$$

すなわち，許容方向の中から QP 問題の目的関数が最も増加する方向を求める．得られた $\mathbf{d} \in \mathbb{R}^{\ell}$ の成分から q 個を選び出す．その方法は以下の通りである．$\omega_j = y_j(-\sum_{i=1}^{\ell} \alpha_i y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + 1)$ を増加順にソートする．q は偶数に取る．ソートされたリストの先頭から，$0 < \alpha_j < C$ をみたすもの，あるいは $\alpha_j = 0$ もしくは $\alpha_j = C$ で $d_j = -y_j$ をみたすものを $q/2$ 個選ぶ．次に，リストの最後から $0 < \alpha_j < C$ をみたすもの，あるいは $\alpha_j = 0$ もしくは $\alpha_j = C$ で $d_j = y_j$ をみたすものを $q/2$ 個選ぶ．これらの q 個を working set とする．

その他に，SVM^{light} では (24) を速く解くために様々なヒューリスティックスを工夫している．詳しくは [5] を参照されたい．また，収束性の議論が [9] でなされている．

5.3.2 SMO(Sequential Minimal Optimization)

SMO は上記の変数集合 B の構成要素の数を 2 に固定して QP を解析的に解いてしまう．working set に選ばれた変数を α_1, α_2 とする．α_1, α_2 の現在の値を $\alpha_1^{old}, \alpha_2^{old}$ とすると，等式制約条件 $\sum_{i=1}^{\ell} \alpha_i y_i = 0$ がみたされるためには $0 \leq \alpha_1, \alpha_2 \leq C$ の制約の下で

$$\alpha_1 y_1 + \alpha_2 y_2 = \alpha_1^{old} y_1 + \alpha_2^{old} y_2 = \text{定数}$$

となる必要がある．この制約の下に 2 変数の 2 次関数の最大化問題を解析的に解くことはそれほど困難なことではない．部分問題が 2 変数の問題であるので，反復が数多く行われることは容易に予想される．したがって SMO の効率は集合 B をどのように選ぶかということに大きく依存する．ここにもいくつかのヒューリスティックスが利用されてアルゴリズムの効率化に寄与している．[12] を参照されたい．

6 関連した定式化とアルゴリズム

本節では，上で述べたサポートベクターマシンの標準的な定式化とは異なった，しかし類似の方法について簡単に解説する．

6.1 線形計画法を利用した識別

Mangasarian は上述のサポートベクターマシンとは別に数理計画を利用した識別法を提案している．初期の方法は超平面で分離する考えは同じであるが，マージン最大化の考えは入っていない．誤った識別をなるべく少なくする分離ということで，以下のように定式化される．

$$\begin{array}{ll} \text{最小化} & \sum_{i=1}^{\ell} \xi_i, \qquad\qquad\qquad\qquad \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^{\ell} \\ \text{条　件} & y_i \left((\mathbf{w} \cdot \mathbf{x}_i) + b \right) \geq 1 - \xi_i, \xi_i \geq 0, \quad i = 1, ...\ell \end{array}$$

この定式化ではマージン最大化を考慮していないので，上述の統計的性質は期待できない．また，線形分離可能なデータセット（$\boldsymbol{\xi}^* = 0$）に対しては一意の解を与えない．しかし，この問題は線形計画問題なので，かなり大規模なものでも直接解くことが可能であるというメリットがある．

そこで，ソフトマージン最大化を一般化して以下のように定式化してみる：

$$\begin{array}{ll} \text{最小化} & \|\mathbf{w}\|_p + C\sum_{i=1}^{\ell}\xi_i, \qquad\qquad \mathbf{w}\in\mathbb{R}^N, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^\ell \\ \text{条　件} & y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)\geq 1-\xi_i, \xi_i\geq 0 \quad i=1,...\ell \end{array}$$

ここで，$\|\cdot\|_p$ は L_p ノルムを表す．この問題は一般に（凸）非線形計画問題となり，何らかの識別法を与えるであろう．2 次計画問題による定式化との類似から $\mathbf{w}=\sum_{i=1}^{\ell}\alpha_i y_i\mathbf{x}_i$ とおくと，

$$\begin{array}{ll} \text{最小化} & \left\|\sum_{i=1}^{\ell}\alpha_i y_i\mathbf{x}_i\right\|_p + C\sum_{i=1}^{\ell}\xi_i, \qquad\qquad \boldsymbol{\alpha}\in\mathbb{R}^\ell, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^\ell \\ \text{条　件} & y_i\left(\sum_{i,j=1}^{\ell}\alpha_j y_i(\mathbf{x}_i\cdot\mathbf{x}_j)+b\right)\geq 1-\xi_i,\ \xi_i\geq 0, \quad i=1,...\ell \end{array}$$

となるが，目的関数をさらに簡単化して，

$$\begin{array}{ll} \text{最小化} & \|\boldsymbol{\alpha}\|_1 + C\sum_{i=1}^{\ell}\xi_i, \qquad\qquad \boldsymbol{\alpha}\in\mathbb{R}^\ell, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^\ell \\ \text{条　件} & y_i\left(\sum_{i,j=1}^{\ell}\alpha_j y_i(\mathbf{x}_i\cdot\mathbf{x}_j)+b\right)\geq 1-\xi_i,\ \xi_i\geq 0, \quad i=1,...\ell \end{array}$$

とする．カーネル関数を導入すると

$$\begin{array}{ll} \text{最小化} & \|\boldsymbol{\alpha}\|_1 + C\sum_{i=1}^{\ell}\xi_i, \qquad\qquad \boldsymbol{\alpha}\in\mathbb{R}^\ell, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^\ell \\ \text{条　件} & y_i\left(\sum_{i,j=1}^{\ell}\alpha_j y_i K(\mathbf{x}_i,\mathbf{x}_j)+b\right)\geq 1-\xi_i,\ \xi_i\geq 0, \quad i=1,...\ell \end{array}$$

を得る．このような定式化ではマージン最大化ではなくて，スパース性を追求したことになることは明らかであろう．この問題は，補助変数を導入することによって線形計画問題となるので，リニアプログラミングマシンによる識別と呼ばれている．識別関数は (21) と同様である．

6.2　半正定値計画法を利用した識別

超平面による識別法を拡張して，2 次曲面による分離法も提案されている ([8])．すなわち，超平面

$$(\mathbf{w}\cdot\mathbf{x})+b=0$$

の代わりに，2 次曲面

$$(\mathbf{x}\cdot\mathbf{D}\mathbf{x})+(\mathbf{w}\cdot\mathbf{x})+b=0$$

を考える．2 次曲面の自由度を減らすために，係数行列 $\mathbf{D}$ を半正定値行列に限定する．したがって，問題は

$$\begin{array}{lll} \text{最小化} & \sum_{i=1}^{\ell}\xi_i, & \mathbf{D}\in\mathbb{R}^{N\times N}, \mathbf{w}\in\mathbb{R}^N, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^\ell \\ \text{条　件} & y_i\left((\mathbf{x}_i\cdot\mathbf{D}\mathbf{x}_i)+(\mathbf{w}\cdot\mathbf{x}_i)+b\right)\geq 1-\xi_i,\ \xi_i\geq 0, & i=1,...\ell \\ & \mathbf{D}\succeq\mathbf{0} & \end{array}$$

と定式化される．この問題は，半正定値計画問題と呼ばれる問題である．今野たちは切除平面による実用解法を提案して，倒産判別問題に応用している ([7])．

7　サポートベクターマシンによる倒産判別への応用

SVM の応用例の一つとして，企業の財務データによる倒産・非倒産の判別に適用した例について報告する．この問題は,「信用リスクの計量化と管理」と呼ばれる分野の重要なコンポーネントをなすものである．

倒産企業，非倒産企業それぞれ 250 社の財務データから 16 個の指標を選択して（$[0,1]$ の区間に値を正規化して）学習をおこなった．このような財務データは粉飾されていることが多いが，実用上は粉飾決算データを前提にして判定したいので，あえてそのようなデータを基に解析をおこなっている．本実験では，トレーニングデータがそれほど大規模ではないので，汎用数理計画パッケージ NUOPT の QP ソルバーを利用した．上で述べたように，内点法はサポートベクターの判定に適していないので，主有効制約法を使用した．

(1) 1 ノルムソフトマージン最大化による学習

上で述べた 3 種類のカーネルに対して，ペナルティパラメータ C の値を変化させて実験を行った．学習結果，検証と正答率 (%) を記してある．（*印は各列の最良値を，x 印は各列の最低値を与えるもの.）

カーネル	C	倒産（学習）	非倒産（学習）	倒産（検証）	非倒産（検証）
多項式	10	80.8	81.2	81.3	71.2*
多項式	100	88.0	84.4	75.8	69.4
多項式	1000	90.0	85.6	77.3	67.1
RBF	10	82.8	77.6	83.8	69.8
RBF	100	85.6	82.8	81.3	70.1
RBF	1000	92.4*	87.2*	79.8	66.7
シグモイド	10	66.8x	67.2	73.2	62.9
シグモイド	100	67.6	69.2	71.7	63.3
シグモイド	1000	68.0	68.4	71.2x	63.1
線形	10	87.2	65.6x	90.9*	60.9x

計算に使用したパラメータ値は以下のとおり:多項式カーネルの $d=2$, RBF カーネルの $2\sigma^2=3.0$, シグモイドカーネルの $\kappa=1, \theta=3.78$.

計算結果から以下のような事実が観測される．

(i) 学習結果は，ペナルティパラメータの値が大きい方が良くなり，値が小さい方が悪くなるがるが，検証結果は逆にペナルティパラメータの値が小さい方が良好である．これは，C の値が大きすぎると過剰学習（汎化能力の低下）が起こっていることを示している．

(ii) 倒産判別に関しては，多項式カーネルと RBF カーネルはほぼ同等の性能を示す．シグモイドカーネルはあまり良くない．単純な線形分離（非線形カーネルを用いないもの）でも，ある程度良好な結果を与えている．

(2) 2 ノルムソフトマージン最大化による学習

2 ノルムを用いた場合を以下に示す．$C=10$ と固定した．計算に使用したパラメータ値は 1 ノルムの場合と同様である．

カーネル	倒産（学習）	非倒産（学習）	倒産（検証）	非倒産（検証）
多項式	69.9x	81.2*	80.8x	69.9
RBF	81.6	79.2	82.3	70.6*
線形	82.0*	71.2x	86.4*	69.0x

この場合も 1 ノルムの場合と同様に多項式カーネルと RBF カーネルの成績はほぼ同等である．線形分離は検証結果において良い成績を出していることが注目される．

参考文献

[1] N.Cristianini and J.Shawe-Talor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[2] O. L. Mangasarian, Linear and nonlinear separation of patterns by linear programming, *Operations Research*, 13 (1965), pp.444-452.

[3] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers, in *5th Annual ACM Workshop on COLT*, D. Haussler, ed., pp.144-152, ACM Press, 1992.

[4] P. Bradley and O. Mangasarian, Massive data discrimination via linear suppport vector machines, Mathematical Programming Technical Report 98-05, University of Wisconsin Madison, 1998.

[5] T. Joachims, Making large-scale SVM learning practical, in B. Schölkopf, C. J. C. Burges, and A. J. Smola, edits, *Advances in Kernel Methods — Support Vector Learning*, pp. 169-184, Cambridge, MA, MIT Press, 1999.

[6] L. Kaufman, Solving the quadratic programming problme arising in support vector classification, in B. Schölkopf, C. J. C. Burges, and A. J. Smola, edits, *Advances in Kernel Methods — Support Vector Learning*, pp. 147-167, Cambridge, MA, MIT Press, 1999.

[7] H. Konno, J. Gotoh, and T. Uno, A cutting plane algorithm for semi-definite programming problems and applications to failure discrimination and cancer diagnosis, CRAFT WP 00-07, Center for Research in Advanced Financial Technology, Tokyo Institute of Technology, 2000.

[8] H. Konno, and H. Kobayashi, Failure discrimination and rating of enterprises by semi-definite programming, *Asia Pacific Financial Markets*, 7, pp.261-273, 2000.

[9] Chih-Jen Lin, On the convergence of the decomposition methods for support vector machines. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2000.

[10] M. Oren, C. Papageorgiou, P. Sinha, and E. Osuna, Pedestrian detection using wavelet templates, in *Proceedings of CVPR'97*, Puerto Rico, 1997.

[11] E. Osuna, R. Freund, and F. Girosi, Training support vector machines: An application to face detection, in *Proceedings of CVPR'97*, Puerto Rico, 1997.

[12] J. Platt. Fast training of support vector machines using sequential minimal optimization, in *Advances in Kernel Methods — Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, eds, pp. 185-208, Cambridge, MA, MIT Press, 1999.

[13] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods — Support Vector Learning*, MIT Press, Cambridge, MA, 1999.

[14] A. J. Smola, P. L. Barlett, B. Schölkopf, and D.Schuurmans, *Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA, 2000.

[15] V.N.Vapnik, *Statistical Learning Theory*, J.Wiley, 1998.

[16] V.N.Vapnik, *The Nature of Statistical Learning Theory* (Second Edition), Springer, 1999.

[17] V. Vapnik and A. Chervonenkis, A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.