

サポートベクトルマシンとカーネル法

高野 祐一

サポートベクトルマシンは最適分離超平面に基づくパターン認識手法であり、非線形データ解析の手法であるカーネル法と組み合わせて1992年に提案され、大きな注目を集めた。本稿では、サポートベクトルマシンによる分類と回帰の手法を解説し、さらにカーネル法の概要と応用を紹介する。他の手法との関係性を見ながらカーネル法の理解を促進し、さまざまな応用例を通してカーネル法の汎用性を吟味する。

キーワード：サポートベクトルマシン、カーネル法、表現定理、 K 最近傍法、深層学習、多変量解析

1. はじめに

パターン認識では、認識対象について観測可能な情報（説明変数）から、その対象のクラス（目的変数）を識別する規則を学習する。パターン認識は人工知能の基盤技術としてさまざまな場面で利用される（表1）。銀行の融資審査の例では、融資先の企業が倒産してしまうと銀行は資金を回収することができなくなる。したがって企業の財務指標などの情報を用いて、将来倒産する企業を予測することができれば有用である。

サポートベクトルマシン（support vector machine: SVM）は最適分離超平面に基づくパターン認識手法である。非線形データ解析の手法であるカーネル法 [1] と組み合わせて1992年に提案された [2]。最適分離超平面やカーネル法のアイデアは古くから知られていたが [3, 4]、それらの重要性は十分に認識されておらず、SVMの提案によって脚光を浴びることとなった。文献 [5] でソフトマージン型のSVMが考案され、文献 [6] でSVMは回帰分析へと拡張された。カーネル法は主成分分析・クラスター分析・外れ値の検出など、さまざまな多変量解析法に応用可能な汎用性の高い手法である [4, 7-9]。また近年では、動的ポートフォリオ選択にもカーネル法が利用されている [10, 11]。

本稿では、SVMによる分類と回帰の手法を解説し、カーネル法の概要と応用を紹介する。近年は深層学習（多層の人工ニューラルネットワーク）が盛んに研究されており、その影響もあってカーネル法の存在感が薄くなっているように感じられる。しかしながら深層学習と比較して、カーネル法には以下のような利点があると考える。

- 訓練事例や計算資源が少ない状況でも、十分な性

表1 パターン認識の例

例	対象	説明変数	目的変数
文字認識	手書き文字	画像データ	文字
音声認識	発話音声	音声データ	文字列
病気の診断	患者	問診結果	病名
融資審査	企業	財務指標	存続/倒産



図1 パターン認識のイメージ

能を発揮できる。

- 豊富な数学理論（理論保証）が整備されている。

本稿の主な目的は、カーネル法の本質を理解し、カーネル法の価値を見直す機会を読者に提供することである。SVMやカーネル法に興味をもった読者は、5節で紹介する文献を読んでさらに理解を深めてほしい。

2. サポートベクトルマシン

本節では、SVMによる分類（SV分類）と回帰（SV回帰）の手法を説明し、企業格付の分析例を通してSVMの有用性を確認する。

SVMなどのパターン認識手法では、 p 個の説明変数から成るベクトル $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$ を入力し、目的変数 y を正しく出力するような識別器（分類規則）を学習する（図1）。

2.1 ハードマージンSV分類

2クラス分類問題を想定し、目的変数は2値ラベル $y \in \{-1, +1\}$ とする。切片 b と係数ベクトル $\mathbf{w} = (w_1, w_2, \dots, w_p)^T$ を導入し、線形識別関数を以下のように定義する。

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{j=1}^p w_j x_j + b \quad (1)$$

たかの ゆういち
筑波大学システム情報系
〒305-8573 茨城県つくば市天王台 1-1-1

この関数の出力値の符号に基づく以下の識別器を考える。

$$\begin{cases} f(\mathbf{x}) < 0 & \Rightarrow y = -1 \\ f(\mathbf{x}) \geq 0 & \Rightarrow y = +1 \end{cases}$$

識別器の学習に使用する n 個の訓練事例の集合を $\{(\mathbf{x}_i, y_i) \mid i \in [n]\}$ とする。ここで、 $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^\top$ は事例 i の説明変数のデータ、 $y_i \in \{-1, +1\}$ は事例 i の目的変数 (2 値ラベル) のデータとし、 $[n] = \{1, 2, \dots, n\}$ は事例の添え字集合とする。ハードマージン SV 分類では、すべての訓練事例が線形識別関数 (1) により正しく分類できること (線形分離可能条件) を仮定する。すなわち、

$$\begin{cases} y_i = -1 & \Rightarrow \mathbf{w}^\top \mathbf{x}_i + b < 0 \\ y_i = +1 & \Rightarrow \mathbf{w}^\top \mathbf{x}_i + b > 0 \end{cases} \quad (i \in [n]) \quad (2)$$

が成立するような (b, \mathbf{w}) が存在することを仮定する。

線形分離可能条件 (2) の下では、すべての訓練事例を正しく分類できる線形識別関数 (1) は無数に存在する。ハードマージン SV 分類では、未知の事例も正しく分類できるように、マージン (分類境界 $f(\mathbf{x}) = 0$ から最も近い訓練事例までの距離) が最大となるように識別関数 (分類境界) を学習する (図 2)。

ベクトルの L_2 ノルム

$$\|\mathbf{w}\|_2 = \sqrt{w_1^2 + w_2^2 + \dots + w_p^2}$$

を用いて、ハードマージン SV 分類は以下の最適化問題として記述される。

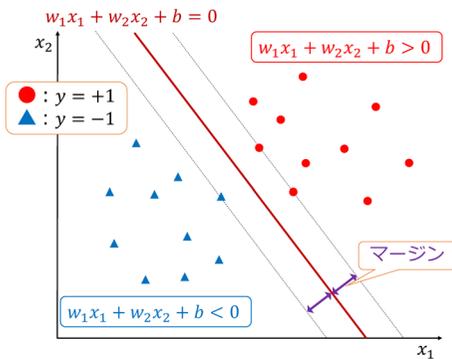


図 2 ハードマージン SV 分類 ($p = 2$)

$$\text{目的関数: } \frac{1}{2} \|\mathbf{w}\|_2^2 \rightarrow \text{最小} \quad (3)$$

$$\text{制約条件: } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad (i \in [n]) \quad (4)$$

$$(b, \mathbf{w}) \in \mathbb{R} \times \mathbb{R}^p \quad (5)$$

式 (4) は式 (2) に相当し、訓練事例がすべて正しく分類されることを表す。式 (5) は最適化問題の決定変数を表す。

式 (3) はマージンの最大化を表すが、このことを直感的に理解するために、説明変数の数を $p = 2$ として高校数学の公式を利用する。点 (x_{i1}, x_{i2}) と直線 $w_1x_1 + w_2x_2 + b = 0$ の距離は以下のように表される。

$$\frac{|w_1x_{i1} + w_2x_{i2} + b|}{\sqrt{w_1^2 + w_2^2}} \quad (6)$$

分数 (6) に対して、式 (4) により分子を一定値以上に抑えて、式 (3) により分母 (の 2 乗) を最小化することで、式 (6) の最小値 (マージン) が最大化されている。

2.2 ソフトマージン SV 分類

一般には、線形識別関数 (1) を用いてすべての訓練事例を正しく分類できるような都合の良い状況は少なく、その場合は線形分離可能条件 (2) を満たすような (b, \mathbf{w}) は存在しない。すなわち式 (4) を成立させることは不可能なので、ソフトマージン SV 分類では分類の違反度を表す非負決定変数 $\xi = (\xi_1, \xi_2, \dots, \xi_n)^\top$ を導入し、なるべく分類の誤りが小さくなるように分類境界を学習する (図 3)。

分類の違反度最小化の優先度を表す非負パラメータ C を導入し、ソフトマージン SV 分類は以下の最適化問題として記述される。

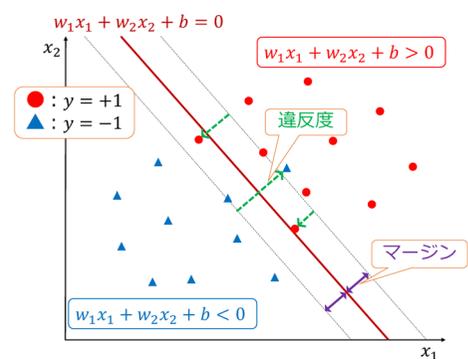


図 3 ソフトマージン SV 分類 ($p = 2$)

$$\text{目的関数: } \underbrace{\frac{1}{2}\|\mathbf{w}\|_2^2}_{\text{正則化項}} + C \underbrace{\sum_{i=1}^n \xi_i}_{\text{損失関数}} \rightarrow \text{最小} \quad (7)$$

$$\text{制約条件: } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \quad (i \in [n]) \quad (8)$$

$$(b, \mathbf{w}, \boldsymbol{\xi}) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}_+^n \quad (9)$$

式 (8) で事例 i の分類の違反度 ξ_i が計上され、式 (7) で違反度の総和が最小化されている。

式 (7) は正則化項（モデルの複雑さに対する罰則）と損失関数（訓練事例に対する誤りの大きさ）の和と解釈することができる。損失関数の最小化のみに注力すると、訓練事例に過剰に適合した識別関数が学習されてしまい、未知の事例に対する予測能力（汎化性能）が却って悪化してしまう。正則化項は識別関数の係数の大きさに対応し、不必要な係数を縮小させて訓練事例に対する過剰適合を抑制する働きがある。正則化項として L_1 ノルム

$$\|\mathbf{w}\|_1 = |w_1| + |w_2| + \dots + |w_p|$$

を使用すると、係数としてゼロが推定されやすくなり、変数選択の役割も兼ねることができる。

2.3 SV 回帰

回帰分析では、実数値の目的変数 $y \in \mathbb{R}$ に対して、

$$y \approx \mathbf{w}^\top \mathbf{x} + b$$

の関係が成立するように (b, \mathbf{w}) を推定する。最小二乗法では、事例 i の残差 $e_i = y_i - (\mathbf{w}^\top \mathbf{x}_i + b)$ の二乗和が最小となるように推定するが、SV 回帰では非負パラメータ ε を事前に設定し、 $-\varepsilon \leq e_i \leq \varepsilon$ の範囲を超えるような大きな残差 e_i に対してのみ違反度 ξ_i を計上する (図 4)。

SV 回帰は以下の最適化問題として記述される。

$$\text{目的関数: } \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^n \xi_i \rightarrow \text{最小} \quad (10)$$

$$\text{制約条件: } -\varepsilon - \xi_i \leq y_i - (\mathbf{w}^\top \mathbf{x}_i + b) \leq \varepsilon + \xi_i \quad (i \in [n]) \quad (11)$$

$$(b, \mathbf{w}, \boldsymbol{\xi}) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}_+^n \quad (12)$$

式 (11) で範囲 $[-\varepsilon, \varepsilon]$ からの残差の超過量が違反度 ξ_i として計上されている。最小二乗法と比較すると、SV 回帰は小さな残差は損失に加え、損失の増え方が緩やか（2 次関数ではなく 1 次関数）であるため、外れ値による悪影響が小さくなる。

2.4 企業格付の予測

SVM の有用性を確認するために、企業格付の分析

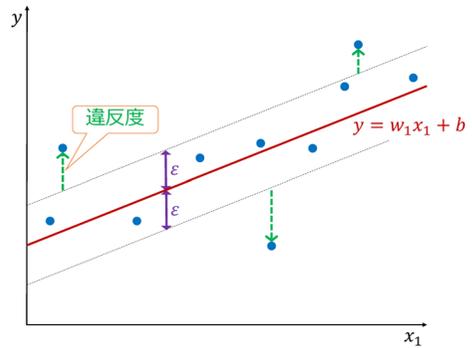


図 4 SV 回帰 ($p = 1$)

表 2 企業格付予測の正答率 [12]

	訓練事例 (2011 年)		
	順序 ロジット	SV 分類 一様	加重
全企業	90.4%	85.2%	83.7%
下位企業	83.3%	70.8%	85.0%
上位企業	93.3%	91.2%	83.2%
	検証事例 (2012 年)		
	順序 ロジット	SV 分類 一様	加重
全企業	47.3%	84.7%	82.7%
下位企業	29.8%	70.2%	83.9%
上位企業	55.4%	91.4%	82.2%

例 [12] を紹介する。51 種類の財務指標を説明変数として、上位企業（格付 A- 以上：約 120 社）と下位企業（格付 A- より下：約 280 社）を分類する。2011 年のデータを訓練事例として識別関数を学習し、2012 年のデータを検証事例として予測性能を評価する。

順序ロジットモデルと SV 分類による企業格付予測の正答率を表 2 に示す。順序ロジットモデルは順序選択のための統計的モデルあり、ここでは SV 分類は金融リスク指標に基づく拡張がなされており、加重型の SV 分類では各クラスの事例の重みが調整されている [12]。順序ロジットモデルは訓練事例に対しては高い正答率を示すが、検証事例に対する正答率は大きく低下している。一方で SV 分類は正則化項の効果により、検証事例に対しても高い正答率を維持している。

3. カーネル法の概要

本節では非線形データ解析の手法であるカーネル法について説明し、他の手法との関係性を見ながらカーネル法の理解を促進する。

3.1 特徴写像

線形の分類境界では適切な分類ができない場合を考

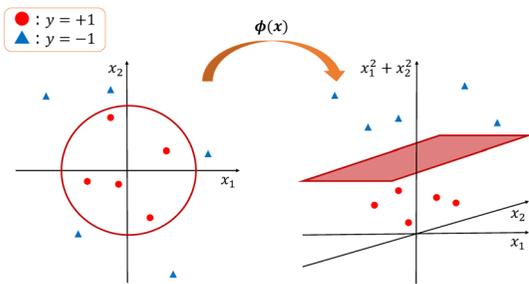


図5 特徴写像による非線形分類 ($p = 2$)

える。すなわち図5左の円周のような非線形分類境界を学習するために、説明変数の非線形項(特徴量)のベクトルを表す特徴写像 $\phi(\mathbf{x})$ を導入する。ここでは特徴量 $x_1^2 + x_2^2$ を加えた3次元空間への特徴写像を考える。

$$\phi(\mathbf{x}) = (x_1, x_2, x_1^2 + x_2^2)^\top \quad (13)$$

この特徴写像を用いた識別関数は以下のようにになる。

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w}^\top \phi(\mathbf{x}) + b \\ &= w_1 x_1 + w_2 x_2 + w_3 (x_1^2 + x_2^2) + b \end{aligned}$$

特徴写像(13)に基づく3次元の特徴空間(図5右)では、円の外側の点(図5左)が上方に布置される。それゆえ線形分類境界(平面)でクラスを分類することができ、図5左では非線形分類境界(円周)に相当する。

この例では特徴量 $x_1^2 + x_2^2$ が有効に働いたが、実用上はどのような特徴量が有効か事前にはわからない。そのため多種多様な非線形項を用意せざるをえなくなり、特徴空間の次元が上がって計算量が増大してしまう。

3.2 表現定理

カーネル法では、上述の問題点を巧みな方法で解決する。まずは、特徴空間上の内積を表すカーネル関数を導入しよう。

$$k(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_2) \quad (14)$$

二つのベクトルは方向が同じ場合に最も内積が大きくなるので、カーネル関数は事例間の(特徴空間上の)類似度を表すものと解釈できる。ただし、特徴空間の次元が高いと内積(14)を計算することも難しいので、カーネル法では高次元空間上の内積としての性質を有する以下のような関数を利用する。

多項式カーネル:

$$k(\mathbf{x}_1, \mathbf{x}_2) = (\alpha \mathbf{x}_1^\top \mathbf{x}_2 + \beta)^\delta \quad (15)$$

ガウスクーネル:

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2}{\sigma^2}\right) \quad (16)$$

シグモイドカーネル:

$$k(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{1 + \exp(-\gamma \mathbf{x}_1^\top \mathbf{x}_2)} \quad (17)$$

ここで、 $\alpha, \beta, \delta, \sigma, \gamma$ はカーネル関数のパラメータである。

SV分類やSV回帰では、正則化項と損失関数の和(7), (10)を最小化する。これらの手法で特徴写像 $\phi(\mathbf{x})$ を利用するとき、以下の定理が成り立つ。

定理1(表現定理[7])。正則化項(L_2 ノルム)と損失関数の和を最小化の問題に対して、最適識別関数 $f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b$ は、カーネル関数の線形和

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b \quad (18)$$

の形式で書ける。

この表現定理とカーネル関数(15)–(17)を利用すれば、特徴写像を陽に扱うことなく、高次元空間上の計算を回避して最適識別関数(18)を学習でき、非線形分類や回帰が実現できる。定理の証明は付録Aに記載することとし、次節では他の手法との関係性を通して表現定理の理解を深める。

3.3 K最近傍法との関係

K最近傍法では、識別対象と類似するK個の事例を選定し、それらの平均値(多数決)に着目する。

$$f(\mathbf{x}) = \frac{1}{K} \sum_{i \in N_K(\mathbf{x})} y_i \quad (19)$$

ただし、 $N_K(\mathbf{x})$ は対象事例 \mathbf{x} と類似するK個の訓練事例の集合(近傍)とする。たとえば $K = 3$ とし、近傍事例の2値ラベル y_i の値が $-1, +1, +1$ である場合は、識別関数の値は

$$f(\mathbf{x}) = \frac{-1 + 1 + 1}{3} = \frac{1}{3} > 0$$

となるため、対象事例 \mathbf{x} のラベルは $y = 1$ と識別する。

ここで近傍 $N_K(\mathbf{x})$ に基づいて、カーネル関数を以下のように定義する。

$$k(\mathbf{x}_i, \mathbf{x}) = \begin{cases} 1/K & i \in N_K(\mathbf{x}) \\ 0 & \text{それ以外} \end{cases}$$

さらに、カーネル関数の係数を目的変数と一致させて $\alpha_i = y_i$ とし、切片を $b = 0$ とする。このとき、カーネル法による識別関数 (18) は K 最近傍法による識別関数 (19) と一致する。この意味でカーネル法は、近傍の事例を集めて多数決を採用する K 最近傍法の一般化 (高機能版) と解釈することができる。

3.4 2層 ANN との関係

2層の人工ニューラルネットワーク (artificial neural network: ANN) を考える (図 6)。ここでは簡単のために説明変数の数は $p = 2$ とし、最初に識別対象の説明変数 $\mathbf{x} = (x_1, x_2)^\top$ が入力される。中間層のユニットは n 個用意されており、訓練事例の説明変数のデータを重みとした総入力

$$u_i = \mathbf{x}_i^\top \mathbf{x} = x_{i1}x_1 + x_{i2}x_2$$

を受け取る。活性化関数 g として以下のシグモイド関数を用いると、中間層のユニットの出力は $\gamma = 1$ のシグモイドカーネル (17) と一致する。

$$\begin{aligned} g(u_i) &= \frac{1}{1 + \exp(-u_i)} \\ &= \frac{1}{1 + \exp(-\mathbf{x}_i^\top \mathbf{x})} = k(\mathbf{x}_i, \mathbf{x}) \end{aligned}$$

係数 α_i を重みとして出力層へ送ると、最終的な出力は

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i g(u_i) + b = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b$$

となり、カーネル法による識別関数 (18) と一致する。この意味でカーネル法は、特殊な形式の 2 層 ANN とみなすことができる。層数を増やして表現力を高める深層学習とは対照的に、中間層で多数のユニット (訓練事例と同数の特徴量) を用意して表現力を高めることがカーネル法の特徴である。

4. カーネル法の応用

カーネル法は SVM 以外にもさまざまな多変量解析法に応用できる。本節ではそれらの応用例を紹介し、カーネル法の汎用性を吟味する。

4.1 主成分分析

主成分分析は相関の強い変数を合成することで、データの次元を縮約する方法である。通常の主成分分析では、変数間の線形関係しか考慮することができないが、カーネル法を利用すれば非線形な関係を有する変数を合成することができる。またカーネル主成分回帰とよばれる手法では、カーネル主成分分析により作成された合成変数を回帰分析の説明変数として利用する。説

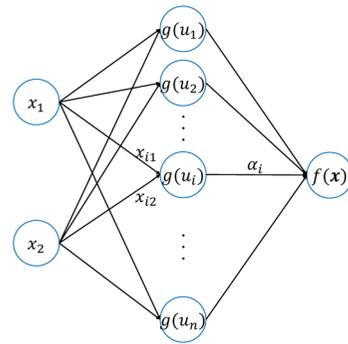


図 6 カーネル法を表す 2 層 ANN ($p = 2$)

明変数が厳選され、ノイズ除去の効果も期待できるため、比較的少ない計算量で汎化性能の高い非線形な識別関数を学習することができる。

4.2 クラスタ分析

クラスタ分析は類似した事例をまとめてグループに分割する手法であり、Ward 法や K 平均法が標準的な方法として知られている。たとえば K 平均法では、グループの代表点と距離が近い事例をグループにまとめていくが、曲線や曲面に沿って分布するような事例のグループを抽出することは難しい。カーネル K 平均法は高次元の特徴空間上で K 平均法を適用し、非線形構造の抽出を可能とする。ただし、 K 平均法には結果が初期値に依存するという問題があるが、固有値問題として定式化されるスペクトラルクラスタリングでは、この問題が緩和される。

4.3 外れ値の検出

データに含まれる外れ値の検出にもカーネル法が利用できる。1 クラス SV 分類とよばれる手法では、多くの訓練事例を包含するようになるべく小さな領域を構成し、その領域から外れた事例を外れ値 (新規性・異常値) として検出する。ここでもカーネル法を利用することで、非線形な領域を構成することが可能となる。

4.4 動的ポートフォリオ選択

ポートフォリオ選択問題では、金融資産への最適な (リスクが低く収益性が高い) 投資比率を決定する。動的な問題設定では状況に応じて柔軟に投資比率を決定するために、その時点までの情報 \mathbf{x} を入力して最適な投資比率を出力する関数 (制御政策) $f(\mathbf{x})$ を推定する。この制御政策の推定にもカーネル法が利用できる [10, 11]。

5. おわりに

本稿では、SVM による分類と回帰の手法を解説し、

カーネル法の概要と応用を紹介した。SVM ではマージンの最大化が正則化項と対応し、汎化性能の向上に寄与している。またカーネル法は表現定理を巧みに利用して非線形データ解析を実現する手法であり、さまざまな多変量解析法や動的意思決定に応用できる。

カーネル法は特殊な形式の 2 層 ANN とみなすことができ、このことから以下のような深層学習との対比が可能であろう。カーネル法は中間層に多数の特徴量を用意し、それらを組み合わせて識別する (図 6)。この意味でカーネル法は、たくさんの料理 (特徴量) を事前に用意しておき、各自が取り分けて食べる「バイキング形式」のような方法といえる。一方で深層学習は、ネットワークを多層化して識別性能が向上するように特徴量を作り込む。この意味で深層学習は、常連客の嗜好に合わせて料理 (特徴量) を提供する「シェフのおまかせ形式」のような方法といえる。どちらの形式が好みかは意見が分かれるところだが、バイキング形式であれば誰でもそれなりに満足できるだろう。カーネル法も同様に多くの状況で有効に機能する万能型の手法であり、一方で深層学習は特定の状況で抜群の性能を発揮する職人芸のような手法といえる。

読者がさらに学習を深めるために、有用な文献を紹介する。SVM については文献 [3, 13, 14] など、カーネル法については文献 [4, 7-9] などがある。筆者は学生時代に文献 [7] を読んでカーネル法に魅了され、その後の研究 [10, 11] にもつながった。

謝辞 本稿の執筆にあたり、有益な助言をいただいた後藤順哉氏 (中央大学)、岩永二郎氏 ((株) エルデシユ)、鮎川矩義氏 (東京理科大学)、西村直樹氏 ((株) リクルートライフスタイル)、小林健氏 ((株) 富士通研究所) に感謝いたします。

参考文献

[1] M. Aizerman, E. Braverman and L. Rozonoer, “Theoretical foundations of the potential function method in pattern recognition learning,” *Automation and Remote Control*, **25**, pp. 821–837, 1964.
 [2] B. E. Boser, I. M. Guyon and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pp. 144–152, 1992.
 [3] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2005. (大北剛訳, 『サポートベクターマシン入門』, 共立出版, 2005.)
 [4] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004. (大北剛訳, 『カーネル法によるパターン解析』, 共立

出版, 2010.)
 [5] C. Cortes and V. N. Vapnik, “Support-vector networks,” *Machine Learning*, **20**, pp. 273–297, 1995.
 [6] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
 [7] 赤穂昭太郎, 『カーネル多変量解析』, 岩波書店, 2008.
 [8] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2001.
 [9] 杉山将, 『イラストで学ぶ機械学習』, 講談社, 2013.
 [10] Y. Takano and J. Gotoh, “A nonlinear control policy using kernel method for dynamic asset allocation,” *Journal of the Operations Research Society of Japan*, **54**, pp. 201–218, 2011.
 [11] Y. Takano and J. Gotoh, “Multi-period portfolio selection using kernel-based control policy with dimensionality reduction,” *Expert Systems with Applications*, **41**, pp. 3901–3914, 2014.
 [12] J. Gotoh, A. Takeda and R. Yamamoto, “Interaction between financial risk measures and machine learning methods,” *Computational Management Science*, **11**, pp. 365–402, 2014.
 [13] 前田英作, “痛快! サポートベクトルマシン—古くて新しいパターン認識手法—,” *情報処理*, **42**, pp. 676–683, 2001.
 [14] 竹内一郎, 鳥山昌幸, 『サポートベクトルマシン』, 講談社, 2015.

付録 A. 定理 1 の証明

訓練事例を入力した特徴画像 $\phi(\mathbf{x}_i)$ によって生成される部分空間を W とし、その直交補空間を W^\perp とする。このとき一般のベクトル \mathbf{w} は、

$$\mathbf{w}_0 = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i) \in W, \quad \boldsymbol{\eta} \in W^\perp$$

なる W と W^\perp の元が存在して、

$$\mathbf{w} = \mathbf{w}_0 + \boldsymbol{\eta}$$

と分解できる。

このとき $\boldsymbol{\eta}^\top \phi(\mathbf{x}_i) = 0$ より、識別関数は

$$f(\mathbf{x}_i) = \mathbf{w}^\top \phi(\mathbf{x}_i) + b = \mathbf{w}_0^\top \phi(\mathbf{x}_i) + b$$

となり、 $\boldsymbol{\eta}$ に依存しない。また $\mathbf{w}_0^\top \boldsymbol{\eta} = 0$ より、正則化項は

$$\frac{1}{2} \|\mathbf{w}\|_2^2 = \frac{1}{2} \mathbf{w}^\top \mathbf{w} = \frac{1}{2} (\|\mathbf{w}_0\|_2^2 + \|\boldsymbol{\eta}\|_2^2)$$

となり、 $\boldsymbol{\eta} = \mathbf{0}$ のときに最小値を取る。したがって、最適解では $\mathbf{w} = \mathbf{w}_0$ が成立し、最適識別関数は以下のように書ける。

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w}_0^\top \phi(\mathbf{x}) + b = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}) + b \\ &= \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b \end{aligned}$$