カーネル法とサポートベクターマシン

池田 盛一*

Kernel Method and Support Vector Machine

Seiichi IKEDA

Abstract – In the nonlinear data analysis, the kernel method is actively done now. Here, the kernel method is described in detail, and Support Vector Machine as the application of this method is shown.

Key words : kernel method, support vector machine, reproducing kernel Hilbert space

1 はじめに

現在, 非線形データへの適用としてカーネル法が注目されている [1]. ここではこのカーネル法の原理や手法について述べ, その適用例としてサポートベクターマシン（SVM：*support vector machine*）を挙げてみる.

2 カーネル法

本節で, カーネル法（*kernel method*）について詳しく述べる.

カーネル法とは, 入力ベクトル $\boldsymbol{x}$ を特徴空間（*feature space*）と呼ばれる高次元の空間の中に $\Phi(\boldsymbol{x})$ として非線形写像し, 識別を行う手法である. 特徴空間での内積の演算は, *Mercer* カーネルと呼ばれる正定値カーネルを用いる. ここで, 次に述べる *Mercer* カーネルの性質から, 写像 Φ の形を陽に知ることなく高次元での計算が可能になる.

定理 1 (Mercer) [2]
入力空間を $\mathcal{X}$ とし, $(\mathcal{X}, \mu)$ を測度有限な測度空間とする. $L_\infty(\mathcal{X})$ を $\mathcal{X}$ 上で定義される有界な可測関数全体の集合とし, $L_2(\mathcal{X})$ を $\mathcal{X}$ 上で定義される 2 乗可積分な可測関数全体の集合とする. $\kappa \in L_\infty(\mathcal{X}^2)$ を対称なカーネル関数とし, 以下の線形変換 $T_\kappa : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ が正定値作用素であるとする.

$$(T_\kappa f)(\cdot) = \int_{\mathcal{X}} \kappa(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) d\mu(\boldsymbol{x}) \qquad (1)$$

このとき, $\psi_i \in L_2(\mathcal{X})$ を T_κ の固有値 $\lambda_i \neq 0$ に対応する正規化固有関数とすると,

1. $(\lambda_1, \lambda_2, \cdots,) \in l_1$
2. $\psi \in L_\infty(\mathcal{X})$ かつ $\sup_i \|\psi_i\|_{L_\infty} < \infty$
3. $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i \in \mathbf{N}} \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{y})$ がほとんどすべての $(\boldsymbol{x}, \boldsymbol{y})$ について絶対一様収束する.

が成り立つ. ここで l_i は絶対和が有界な数列全体の集合である. ■

上の定理は次の形で記述されることもある.

定理 2 (Mercer) [3]
$\mathcal{X}$ は $\mathbf{R}^m$ の有界閉集合で, 関数 $\kappa : \mathcal{X}^2 \to \mathbf{R}$ は連続かつ対称（ $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa(\boldsymbol{y}, \boldsymbol{x})$ ）とする. このとき, 関数 κ が一様収束する級数：

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i \in \mathbf{N}} \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{y}), \quad \lambda_i > 0 \qquad (2)$$

によって展開可能となる必要十分条件は

$$\int_{\mathcal{X}^2} \kappa(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x}) f(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \geq 0, \ \forall f \in L_2(\mathcal{X}) \qquad (3)$$

である. ■

* 釧路高専 一般教科 (数学)

この定理の正定値カーネルの条件 (3) は

$$\sum_{i,j=1}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) c_i c_j \geq 0, \quad \forall c_i, c_j \in \mathbf{R} \tag{4}$$

と表すことができる．すなわち，カーネルグラム行列 (*kernel Gram matrix*)$\{\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j),\ i, j = 1, \cdots, n\}$ と呼ばれる対称行列の各成分が非負定値（半正定値）という条件である．

定理 3 (Moore) [4]
κ が集合 $\mathcal{X}$ 上の正定値カーネルならば，$\mathcal{X}$ 上の関数空間 $\mathcal{H}_\kappa$ が一意に存在して，次の 3 つを満たす．

1. $\kappa(\cdot, \boldsymbol{x}) \in \mathcal{H}_\kappa \qquad (\forall \boldsymbol{x} \in \mathcal{X})$
2. $f = \sum_{i=1}^{n} c_i \kappa(\cdot, \boldsymbol{x}_i)$ は $\mathcal{H}_\kappa$ で稠密．
3. $f(\boldsymbol{x}) = \langle f, \kappa(\cdot, \boldsymbol{x}) \rangle \quad (\forall f \in \mathcal{H}_\kappa, \boldsymbol{x} \in \mathcal{X})$
   但し、内積 $\langle \cdot, \cdot \rangle$ は $\mathcal{H}_\kappa$ での内積とする．■

上記の定理は大変重要で、ここに出てきた $\kappa(\cdot, \boldsymbol{x})$ に対して次の定義を用意しておく．

定義 1 (RKHS と再生核) [5]
集合 $\mathcal{X}$ 上の関数ヒルベルト空間 $\mathcal{H}$ において

1. 任意の$\boldsymbol{x} \in \mathcal{X}$ に対して$\kappa(\cdot, \boldsymbol{x}) \in \mathcal{H}$ となる．
2. $\mathcal{H}$ での内積を $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ で表す．
   任意の $f \in \mathcal{H}, \boldsymbol{x} \in \mathcal{X}$ に対して

   $$\langle f(\cdot), \kappa(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}} = f(\boldsymbol{x}) \tag{5}$$

   を満たす $\mathcal{X} \times \mathcal{X}$ 関数 $\kappa(\cdot, \boldsymbol{x})$ が存在する．

このとき，関数ヒルベルト空間 $\mathcal{H}$ を**再生核ヒルベルト空間**（RKHS：*reproducing kernel Hilbert space*）といい，関数 $\kappa(\cdot, \boldsymbol{x})$ を**再生核**（*reproducing kernel*）という．■

命題 1 [6]
任意の正定値カーネル κ に対して，κ が再生核となるような再生核ヒルベルト空間 $\mathcal{H}$ が存在する．■

命題 2 [5]
再生核 $\Phi(\boldsymbol{x}) = \kappa(\cdot, \boldsymbol{x})$ は，$\phi_x = \kappa(\cdot, \boldsymbol{x})$ とおくことにより，正定値カーネル $\phi_x(\boldsymbol{y}) = \kappa(\boldsymbol{y}, \boldsymbol{x})$ を定める．■

従って，入力空間 $\mathbf{R}^m$ から特徴空間 $\mathcal{F}$ への写像 Φ を

$$\Phi : \mathbf{R}^m \to \mathcal{F} \qquad \boldsymbol{x} \mapsto \kappa(\cdot, \boldsymbol{x})$$

によって定義すれば，

$$\begin{aligned} \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle &= \langle \kappa(\cdot, \boldsymbol{x}), \kappa(\cdot, \boldsymbol{y}) \rangle \\ &= \kappa(\boldsymbol{x}, \boldsymbol{y}) \end{aligned}$$

となり [5]，特徴空間 $\mathcal{F}$ への写像後の内積が入力空間 $\mathbf{R}^m$ での関数として計算が可能になる．つまり，非線形写像によって変換された特徴空間 $\mathcal{F}$ の $\Phi(\boldsymbol{x})$ や $\Phi(\boldsymbol{y})$ を陽に計算する代わりに，$\kappa(\boldsymbol{x}, \boldsymbol{y})$ から最適な非線形写像を構成できることになる．このとき，カーネル関数として入力対象の類似性をうまく表現するものを選んでおけば，特徴空間への写像にユーザーが指定した類似度が反映されることになる．

このように高次元に写像しておきながら，実際には写像された空間での計算を避けて，カーネルの計算のみで最適な識別関数を構成するテクニックのことを「カーネルトリック」と呼ぶ．カーネルトリックは，線形モデルで表される手法を，非線形に拡張する枠組みを示している．

入力空間が連続値の場合（$\mathcal{X} = \mathbf{R}^m$），正定値カーネルの例としては，
identical kernel：

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$$

polynomial kernel：

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + 1)^d$$

gaussian kernel：

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right)$$

sigmoid kernel：

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \tanh(a\langle \boldsymbol{x}, \boldsymbol{y} \rangle - b)$$

などがある．

また，入力空間が離散値の場合にも *convolution kernel* [7] [8]，*Fisher kernel* [9]，*TOP kernel* [10]，*graph kernel* [11] [12] などが提案されている．

3　サポートベクターマシン

サポートベクターマシンは，ニューロンのモデルとして最も単純な線形しきい素子を用いて，2クラスのパターン識別器を構成する手法である．訓練サンプル集合から,「マージン最大化」という基準で線形しきい素子のパラメータを学習する．線形しきい素子は，入力データ $(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ に対し $(\boldsymbol{X}_i \in \mathbf{R}^m,\ \boldsymbol{Y}_i \in \{-1, 1\})$，線形識別関数

$$f_w(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + b, \quad \boldsymbol{w}^\top = (\boldsymbol{a}^\top, b)$$

により，

$$\begin{cases} f_w(\boldsymbol{x}) \geq 0 & \Rightarrow y = 1 \text{（クラス 1）} \\ f_w(\boldsymbol{x}) < 0 & \Rightarrow y = -1 \text{ (クラス } -1\text{）} \end{cases}$$

と判定し，未知の $\boldsymbol{x}$ に対しても正しく答えられるように $f_w(\boldsymbol{x})$ が構成される．ここで，ベクトル $\boldsymbol{a}$ はニューロンモデルのシナプス荷重に対応するパラメータであり，記号 $\top$ は転置を意味する．

3.1　マージン最大化

学習データは線形識別が可能であると仮定する．一般には学習データを分類する線形識別関数は無数にあるが，ここでは次の方針で線形識別関数を構成する．すなわち，もっとも近い訓練サンプルとの余裕をマージンと呼ばれる量（ベクトル $\boldsymbol{a}$ の方向で測った学習データのクラス間の距離）で測り，このマージンが最大となるような識別面（超平面）を求める．このとき超平面の周辺にある少数のサンプルは，あたかも超平面をサポートしているかのように見えるため「サポートベクター」と呼ばれている．

マージンの計算については，$(\boldsymbol{a}^\top, b)$ を定数倍しても識別面は不変なので [13]，制約

$$\begin{cases} \min(\boldsymbol{a}^\top \boldsymbol{X}_i + b) = 1 & \boldsymbol{Y}_i = 1 \text{ のとき} \\ \max(\boldsymbol{a}^\top \boldsymbol{X}_i + b) = -1 & \boldsymbol{Y}_i = -1 \text{ のとき} \end{cases}$$

の下でマージン $\frac{2}{\|\boldsymbol{a}\|}$ の最大化を図ればよい．従って，次のように定式化される．すなわち，

$$\text{制約条件 } \boldsymbol{Y}_i(\boldsymbol{a}^\top \boldsymbol{X}_i + b) \geq 1 \ (\forall i) \tag{6}$$

の下で，

$$\min_{a,b} \|\boldsymbol{a}\|^2 \tag{7}$$

とするパラーメタ $\boldsymbol{w}^\top = (\boldsymbol{a}^\top, b)$ を求めること．これは，いわゆる「2次最適化問題」として扱うことができる [14].

3.2　ソフトマージン

さて，ここまでは，学習データは線形識別が可能であると仮定していた．データ $\boldsymbol{X}_i$ の次元 m が訓練サンプル数よりも大きければこのような仮定は成り立つ．しかし，一般にはそうでない場合も考えられる．そこで，多少の識別誤りは許すように制約条件を少し緩めることにする．これは「ソフトマージン」と呼ばれている．

ソフトマージン法では，いくつかのサンプルが識別面を越えて反対側に入ってしまうことを許す．このときの距離をパラメータ ξ_i (≥ 0) を用いて $\frac{\xi_i}{\|\boldsymbol{a}\|}$ とおけば，その和

$$\sum_{i=1}^{n} \frac{\xi_i}{\|\boldsymbol{a}\|}$$

はなるべく小さい方が望ましい（ξ_i をスラック変数と呼ぶ）．このソフトな条件により，(6),(7) は次のように改められる．

$$\text{制約条件 } \boldsymbol{Y}_i(\boldsymbol{a}^\top \boldsymbol{X}_i + b) \geq 1 - \xi_i,\ (\xi_i \geq 0) \tag{8}$$

$$\min_{a,b} \left[\sum_{i=1}^{n} \Big(1 - \boldsymbol{Y}_i(\underline{\boldsymbol{a}^\top \boldsymbol{X}_i} + b)\Big)_+ + \frac{\lambda}{2} \underline{\|\boldsymbol{a}\|^2} \right] \tag{9}$$

但し，$(z)_+ = \max(z, 0)$ とし，λ は正則化パラメータとする．ここで，下線部分が内積（線形の関数）で表示されていることに注意しておく．

3.3　カーネルによる非線形化

今，入力データ $\boldsymbol{X}_i$ を非線形写像によって変換する．このとき，正定値カーネル κ を用意し，このカーネルにより定まる再生核ヒルベルト空間を H とする．先ほどの最適化問題は下記のようになる．

$$\min_{f \in H, b} \left[\sum_{i=1}^{n} \Big(1 - \boldsymbol{Y}_i(\underline{f}(\boldsymbol{X}_i) + b)\Big)_+ + \frac{\lambda}{2} \underline{\|\boldsymbol{f}\|_H^2} \right] \tag{10}$$

ちょうど，下線部分がこの非線形化により RKHS の元が対応している．このとき，この最適化問題の解は

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \kappa(\boldsymbol{x}, \boldsymbol{X}_i) \tag{11}$$

の形で与えられる（*Representer theorem*） [15]. 従って，先ほどの最適化問題をカーネル κ を用い

た表現にすると，

$$\min_{\alpha,b}\Big[\sum_{i=1}^{n}\Big(1-\boldsymbol{Y}_i\sum_{j=1}^{n}\alpha_j\kappa(\boldsymbol{X}_i,\boldsymbol{X}_j)+b\Big)_+ +\frac{\lambda}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j\kappa(\boldsymbol{X}_i,\boldsymbol{X}_j)\Big] \quad (12)$$

となり，この２次最適化問題に帰着する．

4　その他の例

カーネル法を適用した例としては，カーネル主成分分析 (KPCA) [16]，カーネル判別分析 (KDA) [17]，カーネル正準相関分析 (KCCA)，カーネル独立成分分析 (KICA) [18] などがある．基本的には，線形データ解析アルゴリズムを特徴空間で行うことによって，非線形アルゴリズムが得られることが根底にある．つまり,「内積」を使って表される線形手法（射影 [19]，相関，分散共分散など）なら非線形に拡張が可能ということである．

主成分分析 (PCA) では分散が最大になる方向（部分空間）にデータを射影するのが基本的である．カーネル法を適用した場合，カーネルを設定した中で特徴ベクトルの分散の計算を必要とするが，この計算を陽に行わずにカーネルで代用する．

正準相関分析 (CCA) では２種類の多次元データの相関を探るのが基本的である．カーネル法を適用した場合，同様に特徴空間での相関を最大にする射影方向を求めることになるが，ここの計算でもカーネルを用いる．

非線形アルゴリズムの特徴としては，まず，線形ではとらえられない性質が調べられることが挙げられるが，同時にこの非線形性はカーネルの選び方に強く影響を受けることが知られている．

5　今後の課題

今回，この論文は平成１７年度の内地研究員制度を利用した研究の報告を兼ねている．カーネル法の適用については，データの個数と変量の個数がほぼ等しいことになるので，当然の結果として overfitting の問題が生じる．これを回避するために正則化条件を付与することが考えられているが，正則化項によらずに変数選択をする方法を２００７年９月の「第９回日本中国統計学シンポジウム」で発表する予定である．

参考文献

[1] 福水健次, 公開講座「機械学習の最近の話題」, 統計数理研究所, 2004 年 11 月 26 日.

[2] J.Mercer, *Functions of positive and negative type, and their connection with the theory of integral equations.* Trans. Lond. Phil. Soc. (A), vol.209, pp.415-446, 1909.

[3] J.Shawe-Taylor, Nello Cristianini, *Kernel Methods for Pattern Analysis.* Cambridge Univ Pr (Sd), 2004

[4] Saburou Saitoh, *Theory of reproducing kernels and its applications.* Harlow, UK:Longman Scientific & Technical, 1988.

[5] 梅垣壽郎, 大矢雅則, 日合文雄,『復刊 作用素代数入門』, 共立出版株式会社, 2003.

[6] B.Schölkopf, S.Mike, C.J.C.Burges, P.Knirsch, K.R.Müller, G.Rätsch, A.J.Smola, *Input Space Versus Feature Space in Kernel-Based Methods.* IEEE Transactions on Neural Networks, vol.10, no.5, pp.1000-1017, Sept. 1999.

[7] D.Haussler, *Convolution kernels on discrete structures.* Technical Report UCSC-CRL-99-10, UC Santa Cruz, 1999.

[8] M.Collins, N.Duffy, *Convolution kernels for Natural Language.* Advances in Neulal Information Processing Systems, 14, 2002.

[9] T.Jaakkola, D.Haussler, *Exploiting generative models in discriminative classifiers.* Advances in Neulal Information Processing Systems, 11, pp.487-493, 1999.

[10] K.Tsuda, M.Kawanabe, G.Rätsch, S.Sonnenburg, K.R.Müller, *A new discriminative kernel from probabilistic models.* Advances in Neulal Information Processing Systems, 14, pp.977-984, 2002.

[11] P.Mahé, N.Ueda, T.Akutsu, J.-L.Perret, J.-P.Vert, *Extensions of marginalized graph kernels.* Proc. 21th Intern, Conf. Machine Learning pp.552-559, 2004.

[12] H.Kashima, K.Tsuda, A.Inokuchi, *Marginalized Kernels Between Labeled Graphs.* Proc. 20th Intern, Conf. Machine Learning, 2003.

[13] 津田宏治, 麻生英樹, 村田昇, 『統計科学のフロンティア 6 パターン認識と学習の統計学』, 岩波書店, pp.97-138, 2003.

[14] 栗田多喜夫, 『サポートベクターマシン入門』, 産業技術総合研究所 脳神経情報研究部門, 2002 年 7 月 18 日.

[15] 赤穂昭太郎, 『チュートリアル公演 カーネルマシン』, 信学技報 NC2003-34.

[16] B.Schölkopf, A.Smola,K.R.Müller *Nonlinear component analysis as a kernel eigenvalue ploblem.* Neural,Computation, 10, pp.1299-1319, 1998.

[17] S.Mike, G.Rätsch, J.Weston, B.Schölkopf, K.R.Müller *Fisher discriminant analysis with kernels.* Neural Networks for Signal Processing IX. IEEE, pp.41-48, 1999.

[18] Francis R. Bach, Michael I. Jordan *Kernel Independent Component Analysis.* Jounal of Machine Learning Research, 3, pp.1-48, 2002

[19] 鷲沢嘉一, 山下幸彦, 「カーネル標本空間射影法によるパターン認識」, Workshop on Information-Based Induction Sciences(IBIS 2003), Kyoto, Japan, Nov. 2003.

A 計算例：多項式カーネル

$\mathbf{R}^2$ 上の正定値カーネル

$$\kappa(\boldsymbol{x},\boldsymbol{y}) = (\langle \boldsymbol{x},\boldsymbol{y}\rangle)^2 = (x_1y_1 + x_2y_2)^2$$

が定める再生核ヒルベルト空間 $\mathcal{H}_\kappa$ と写像 $\Phi : \mathbf{R}^2 \to \mathcal{H}_\kappa$ を求めたい [1].

今，$\mathcal{H}$ を $\mathbf{R}^2$ 上の 2 次関数全体とする．

$$f(\boldsymbol{z}) = \alpha_{11}z_1^2 + \alpha_{12}(\sqrt{2}z_1z_2) + \alpha_{22}z_2^2$$

とおくとき，$z_1^2, \sqrt{2}z_1z_2, z_2^2$ を正規直交基底として，内積を以下で定義する．すなわち，

$$f(\boldsymbol{z}) = \alpha_{11}z_1^2 + \alpha_{12}(\sqrt{2}z_1z_2) + \alpha_{22}z_2^2$$

$$g(\boldsymbol{z}) = \beta_{11}z_1^2 + \beta_{12}(\sqrt{2}z_1z_2) + \beta_{22}z_2^2$$

このときの内積は以下の通り．

$$\langle f, g\rangle_{\mathcal{H}} = \alpha_{11}\beta_{11} + \alpha_{12}\beta_{12} + \alpha_{22}\beta_{22}.$$

$\mathcal{H}$ は $\mathbf{R}^3$ と同型になることに注意しておく．

まず，$\mathcal{H} \cong \mathcal{H}_\kappa$ であることを示す：

1. $\kappa(\cdot,\boldsymbol{x}) \in \mathcal{H}$ であること．

   $$\kappa(\boldsymbol{z},\boldsymbol{x}) = \boxed{x_1^2}\cdot z_1^2 + \boxed{\sqrt{2}x_1x_2}\cdot\sqrt{2}z_1z_2 + \boxed{x_2^2}\cdot z_2^2$$

   上の式で，枠付き部分を係数とみなせば良い．

2. $\langle f(\cdot), \kappa(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}} = f(\boldsymbol{x})$ であること．

   任意の $\mathcal{H}$ の元

   $$f(\boldsymbol{z}) = \alpha_{11}z_1^2 + \alpha_{12}(\sqrt{2}z_1z_2) + \alpha_{22}z_2^2$$

   に対し，

   $$\begin{aligned}&\langle f(\cdot), \kappa(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}}\\ &= \alpha_{11}x_1^2 + \alpha_{12}(\sqrt{2}x_1x_2) + \alpha_{22}x_2^2\\ &= f(\boldsymbol{x})\end{aligned}$$

カーネルトリックの実現を確認する：

$$\Phi(\boldsymbol{x}) = \kappa(\cdot,\boldsymbol{x}) \leftrightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top$$

上記は $z_1^2, \sqrt{2}z_1z_2, z_2^2$ を基底とした表現であり，特徴空間での内積は次のようになる．

$$\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y})\rangle_{\mathcal{H}} = \kappa(\boldsymbol{x},\boldsymbol{y}) = (x_1y_1 + x_2y_2)^2$$

このカーネルトリックにより，左辺の 3 次元での内積が右辺では 2 次元の計算で済んでいることが分かる．従って，多項式の次数が高ければ圧倒的に $\kappa(\boldsymbol{x},\boldsymbol{y})$ の計算が有利なのは明らかである．