# サポートベクトルマシンについて

渡辺 一帆

平成 15 年 5 月 23 日

## 1 SVM

サポートベクトルマシン (SVM) は, 分類問題を解くのに用いられる手法である。テキスト分類、画像認識、バイオインフォマティックス等の幅広い問題に応用され、多くの問題で高い認識性能が実現できることが報告されている。ここでは、マージン最大化という基準に従う SVM の学習法を紹介する。

### 1.1 SVM の学習 (線形分離可能なとき)

簡単のために 2 クラスの分類問題を考える。入力 $\mathbf{x} \in R^M$, パラメータ $\mathbf{w} \in R^M, b \in R$ を用いて、

$$g(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{1}$$

とするとき、SVM の出力は $\mathrm{sign}(g(\mathbf{x}))$ で与えられ、超平面 $g(\mathbf{x}) = 0$ の上下で入力 $\mathbf{x}$ を 2 つのクラスに分類する。学習データ $\{(\mathbf{x}_i, y_i), i = 1, \cdots, N, \mathbf{x} \in R^M, y_i \in \{-1, +1\}\}$(出力 y は入力がクラス 1 に属するとき+1、クラス 2 に属するときは-1 とする) が与えられたときに、それぞれのクラスの学習データの中で、分離超平面 $g(\mathbf{x}) = 0$ から最も距離の近いものとの距離を最大化するようにパラメータ $\mathbf{w}, b$ が決定される。$\mathbf{w}, b$ を定数倍しても分離超平面は変化しないので、超平面から最も距離の近い学習データ $\mathbf{x}$ に対して、$g(\mathbf{x}) = 1$ となるように、$\mathbf{w}, b$ を決める。ここで、分離超平面から最も距離の近い学習データをサポートベクトルといい、分離超平面からサポートベクトルまでの距離をマージンという。このとき、パラメータ $\mathbf{w}$ をもつ超平面と $\mathbf{x}$ の距離は、$g(\mathbf{x})/\|\mathbf{w}\|$ なので、この超平面のマージンは $1/\|\mathbf{w}\|$ で表される。よって, マージンを最大化するような、$\mathbf{w}, b$ は以下の最適化問題を解くことによって求められる。

$$\text{制約条件：} \ y_i(\mathbf{w} \cdot \mathbf{x} + b) \geq 1 \ \ (i = 1, \cdots, n)$$

$$\text{最小化：} \ \frac{1}{2}\|\mathbf{w}\|^2 \tag{2}$$

この問題の双対問題を考える[1] 。

[1] 目的関数と制約条件の不等式が凸関数である場合の最適化問題

$$\text{制約条件：} \ h_i(\mathbf{w}) \leq 0 \ \ (i = 1, \cdots, n) \quad \text{最小化：} \ f(\mathbf{w})$$

について、次の定理が成り立つ。ここで、$L(\mathbf{w}, \alpha) = f(\mathbf{w}) + \sum_{i=1}^n \alpha_i h_i(\mathbf{w})$。
(クーン・タッカーの定理)
$\mathbf{w}^*$ が最適解であるための必要十分条件は、

$$\frac{\partial L(\mathbf{w}^*, \alpha^*)}{\partial \mathbf{w}} = \mathbf{0}, \ \ \alpha_i^* h_i(\mathbf{w}^*) = 0, \ \ h_i(\mathbf{w}^*) \leq 0, \ \ \alpha_i^* \geq 0 \ \ (i = 1, \cdots, n)$$

を満たす $\alpha^*$ が存在することである。

またこのとき、

$$L(\mathbf{w}^*, \alpha) \leq L(\mathbf{w}^*, \alpha^*) \leq L(\mathbf{w}, \alpha^*)$$

が成り立つ (($\mathbf{w}^*, \alpha^*$) が $L$ の鞍点である) ことが示される。
よって、$L$ を $\mathbf{w}$ について最小化した

$$\min_{\mathbf{w}} L(\mathbf{w}, \alpha)$$

を、$\alpha$ について最大化する問題として、双対問題が定義され、双対問題の最適解を求めることで主問題の最適解を得ることができる。

$\alpha_i$ をラグランジュの未定乗数とするとき、ラグランジュ関数 $L(\mathbf{w}, b, \alpha)$ は、

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \alpha_i \{1 - y_i g(\mathbf{x}_i)\}$$

となる。

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = \mathbf{0}, \quad \frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

より、

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{3}$$

が得られるので、これを、$L(\mathbf{w}, b, \alpha)$ に代入し、$W(\alpha) = L(\mathbf{w}^*, b^*, \alpha)$ と書くと、以下の双対問題が得られる。

$$\textbf{制約条件：}\ \alpha_i \geq 0\ (i = 1, \cdots, n) \text{ かつ } \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\textbf{最大化：}\ W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{4}$$

このとき、最適なパラメータ $\mathbf{w}^*$ は、

$$\mathbf{w}^* = \sum_{i=1}^{n} y_i \alpha_i^* \mathbf{x}_i \tag{5}$$

と表される。

またクーン・タッカー条件から

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

が成り立つ。上式より $\alpha_i^* > 0$ となるとき、対応するデータ $(\mathbf{x}_i, y_i)$ は $y_i g(\mathbf{x}_i) = 1$ を満たすのでサポートベクトルとなっている。(5) 式より、$\mathbf{w}^*$ は $\alpha_i^* = 0$ となるデータに影響されず、$\alpha_i^* > 0$ に対応するサポートベクトルのみを用いて表すことができる。

## 1.2 SVM の学習 (線形分離不可能なとき)

実際の問題では、データが線形分離可能である場合は少ないので、線形分離不可能な場合に SVM を拡張するために以下の方法が取られている。

**(1) ソフトマージン**

線形分離不可能な場合には、全ての学習データを正しく分類するという制約を満たすことができないので、正の値をもつ変数 $\xi_i (i = 1, \cdots, n)$ を用いて、マージン最大化の問題を以下のように変更する。

$$\textbf{制約条件：}\ y_i g(\mathbf{x}_i) \geq 1 - \xi_i, \quad \xi_i \geq 0 \ \ (i = 1, \cdots, n)$$

$$\textbf{最小化：}\ \frac{1}{2}\|\mathbf{w}\|^2 + c \sum_{i=1}^{n} \xi_i \tag{6}$$

この変更は、$i$ 番目のデータを $\xi_i / \|\mathbf{w}\|$ だけ分離超平面の方向にずらして、線形分離を行なうことに対応する。目的関数の第 2 項は、この修正のための損失で、その係数 $c$ によってこの損失の割合が決められる。この場合の双対問題を考える。非負の未定乗数 $\alpha_i, \beta_i$ を用いて、ラグランジュ関数 $L(\mathbf{w}, b, \alpha, \beta)$ は、

$$L(\mathbf{w}, b, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + c \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i \{1 - \xi_i - y_i g(\mathbf{x}_i)\} - \sum_{i=1}^{n} \beta \xi_i$$

となる。

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = \mathbf{0}, \quad \frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0, \quad \frac{\partial L}{\partial \xi_i} = c - \alpha_i - \beta_i = 0$$

これより、再び

$$\mathbf{w}^* = \sum_{i=1}^{N} y_i \alpha_i^* \mathbf{x}_i$$

と表されるので、$L(\mathbf{w}, b, \alpha, \beta)$ に代入することで、双対問題

$$\textbf{制約条件：}\ 0 \le \alpha_i \le c\ (i = 1, \cdots, n)\ \text{かつ} \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\textbf{最大化：}\ W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{7}$$

が得られる。これは元の双対問題に、$\alpha$ に対する制約を新たに加えたものとなっている。

**(2) カーネル法**

データが線形分離不可能な場合に、入力 $\mathbf{x}$ をさらに高次元の空間へ $\varphi(\mathbf{x})$ で写像し、その空間の中で線形分離をおこなうという手法がとられる。この手法はカーネル法と呼ばれ、$\varphi(\mathbf{x})$ に、ある関数 $K$ に対して、

$$(\varphi(\mathbf{x}) \cdot \varphi(\mathbf{x}_i)) = K(\mathbf{x}, \mathbf{x}_i) \tag{8}$$

となるような関係があれば、カーネル関数 $K$ の値を用いるだけで、(4),(5) 式は

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

とするだけでよく、分離超平面は、(1),(5),(8) 式より

$$g(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x}) + b$$

と求めることができる。このようにカーネル関数を用いることで、実際に高次元空間上での内積を計算することなく、高次元空間上での分類問題を解くことができる。

カーネル関数の例として、

(多項式カーネル ($p$ 次))

$$K(\mathbf{x}, \mathbf{x}_i) = ((\mathbf{x} \cdot \mathbf{x}_i) + 1)^p$$

(ガウシアンカーネル)

$$K(\mathbf{x}, \mathbf{x}_i) = \exp(-\|\mathbf{x} - \mathbf{x}_i\|^2 / c)$$

等がある。

# 参考文献

[1] " An Introduction to Support Vector Machines",N.Cristianini,J.Shawe-Taylor,Cambridge university press,2000