サポートベクターマシン

橋本 玄基

2016/10/21

1 概要

サポートベクターマシン (Support Vector Maschine，以下 SVM) はクラス分類器を作成するための教師あり学習手法の一つである [1]．境界面からデータまでの距離「マージン (Margin)」を最大化するように学習を行うため，汎化性能がトレーニングデータに対して最適になることや，最終的に学習に利用するのは「Support ベクトル (Support Vector)」(後述) となるデータのみのため学習計算量が小さい点などがメリットとしてあげられる．本編では，類似手法であるパーセプトロン (Perceptron) について述べた後，線形な SVM の導出，非線形な SVM の導出，分離不可な問題の解決法の順で解説する．

2 パーセプトロン (Perceptron)

2 クラス分類問題を考える．d 次元の特徴ベクトル $\boldsymbol{x} = (x_1, \cdots, x_d)^T$ をクラス C_1,C_2 に分類するとき，バイアス w_0 と重み $\boldsymbol{w} = (w_1, \cdots, w_d)$ を定義したとき識別関数 $f(\boldsymbol{x})$ は

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 \tag{2.1}$$

と表される．識別境界を $f(\boldsymbol{x}) = 0$ とすれば，識別規則は

$$\begin{cases} f(\boldsymbol{x}) \geq 0 \Rightarrow C_1 \\ f(\boldsymbol{x}) < 0 \Rightarrow C_2 \end{cases} \tag{2.2}$$

となる．クラス C_i の識別関数を

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{0i} \tag{2.3}$$

としたとき $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$ とすれば $g(\boldsymbol{x}) \geq 0$ なら C_1，$g(\boldsymbol{x}) < 0$ なら C_2 と判別できる．

perceptron では学習データを用いて重みを学習する．重みとバイアスを初期化し，学習データを識別関数に入力したとき，誤識別が起きたら

$$\begin{cases} C_1\text{を } C_2\text{と誤識別} \Rightarrow \boldsymbol{w}' = \boldsymbol{w} + \epsilon\boldsymbol{x},\ w_0' = \epsilon\boldsymbol{x} \\ C_2\text{を } C_1\text{と誤識別} \Rightarrow \boldsymbol{w}' = \boldsymbol{w} - \epsilon\boldsymbol{x},\ w_0' = -\epsilon\boldsymbol{x} \end{cases} \tag{2.4}$$

と重みを更新する．ϵ は正の定数である．全てのデータが正しく識別されるまで学習を繰り返す．

3 ハード SVM(Hard SVM)

前章と同様に C_1, C_2 を分類する 2 クラス分類問題について考える．d 次元の特徴ベクトル $\boldsymbol{x} = (x_1, \cdots, x_d)^T$，バイアス w_0 と重み $\boldsymbol{w} = (w_1, \cdots, w_d)^T$ を定義し $\mathbf{x} = (1, \boldsymbol{x}^T)^T$，$\mathbf{w} = (w_0, \boldsymbol{w}^T)^T$ とおく．

3.1 線形な SVM

$\boldsymbol{x}$ を用いて表すと線形識別関数は

$$g(\boldsymbol{x}) = w_0 + \boldsymbol{w}^T \boldsymbol{x} \tag{3.1}$$

と表せる．このとき $g(\boldsymbol{x}) = 0$ の超平面でクラスの境界が存在し，識別境界と呼ばれる．その法線は $\boldsymbol{w}$ である．識別境界に対して一定の距離 (マージンと呼ぶ) 以内には学習データが存在しない用にパラメータ $\mathbf{w}$ を決定することを考える．マージンが最大になる $\mathbf{w}$ を求める手法がサポートベクターマシン (SVM) である．

まずはじめに，重みが $\mathbf{w}$ のとき，あるデータ $\boldsymbol{x}$ と識別境界の距離を求める．$\boldsymbol{x}$ と識別境界の垂線と識別境界の交点を $\boldsymbol{x}^\perp$ とすると

$$\boldsymbol{x} = \boldsymbol{x}^\perp + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \tag{3.2}$$

と表せる．このときの $|r|$ が識別境界と $\boldsymbol{x}$ のユークリッド距離である．$\boldsymbol{x}^\perp$ は識別境界上の点なので

$$g(\boldsymbol{x}^\perp) = w_0 + \boldsymbol{w}^T \boldsymbol{x}^\perp = 0 \tag{3.3}$$

が成立する．式 (3.2) より $\boldsymbol{x}^\perp = \boldsymbol{x} - r(\boldsymbol{w}/\|\boldsymbol{w}\|)$ を式 (3.3) に代入すると

$$\begin{aligned} g(\boldsymbol{x}^\perp) &= w_0 + \boldsymbol{w}^T (x - r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}) \\ &= w_0 + \boldsymbol{w}^T \boldsymbol{x} - r\frac{\boldsymbol{w}^T \boldsymbol{w}}{\|\boldsymbol{w}\|} \\ &= w_0 \boldsymbol{w}^T \boldsymbol{x} - r\frac{\|\boldsymbol{w}\|^2}{\|\boldsymbol{w}\|} \\ &= w_0 + \boldsymbol{w}^T \boldsymbol{x} - r\|\boldsymbol{w}\| \\ &= 0 \end{aligned} \tag{3.4}$$

上式を変形すると

$$r = \frac{w_0 + \boldsymbol{w}^T \boldsymbol{x}}{\|\boldsymbol{w}\|} \tag{3.5}$$

が得られるので求めたい識別境界と $\boldsymbol{x}$ の距離 $|r|$ は

$$|r| = \left| \frac{w_0 + \boldsymbol{w}^T \boldsymbol{x}}{\|\boldsymbol{w}\|} \right| = \frac{|w_0 + \boldsymbol{w}^T \boldsymbol{x}|}{\|\boldsymbol{w}\|} = \frac{|g(\boldsymbol{x})|}{\|\boldsymbol{w}\|} \tag{3.6}$$

である．

ここで，重み $\mathbf{w}$ で全てのデータ $\boldsymbol{x}_i$ $(i = 1, \cdots, n)$ が正しいクラスに分類できているとする．この時すべての i に対して

$$\begin{cases} \boldsymbol{x}_i \in C_1 \Rightarrow g(\boldsymbol{x}_i) > 0 \\ \boldsymbol{x}_i \in C_2 \Rightarrow g(\boldsymbol{x}_i) < 0 \end{cases} \tag{3.7}$$

が成り立たなければならない（SVM では $g(\boldsymbol{x}) = 0$ にデータは存在しないと考える）．データ $\boldsymbol{x}_i$ の所属クラスを $y_i \in \{1, -1\}$($y_i = 1$ なら $\boldsymbol{x}_i \in C_1$，$y_i = -1$ なら $\boldsymbol{x}_i \in C_2$) と表現すると

$$y_i g(\boldsymbol{x}_i) > 0, \quad \forall i = 1, \cdots, n \tag{3.8}$$

は式 (3.7) と等価になる．このときあるデータの識別境界までのユークリッド距離は式 (3.6)，(3.8) から

$$\frac{|g(\boldsymbol{x})|}{\|\boldsymbol{w}\|} = \frac{y_i g(\boldsymbol{x})}{\|\boldsymbol{w}\|} \tag{3.9}$$

と表せる．このとき重み $\mathbf{w}$ に対するマージンは

$$\begin{aligned} \min_{i \in \{1, \cdots, n\}} \frac{y_i g(\boldsymbol{x}_i)}{\|\boldsymbol{w}\|} &= \frac{1}{\|\boldsymbol{w}\|} \min_{i \in \{1, \cdots, n\}} y_i g(\boldsymbol{x}_i) \\ &= \frac{1}{\|\boldsymbol{w}\|} \min_{i \in \{1, \cdots, n\}} y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) \end{aligned} \tag{3.10}$$

で定義される．マージンが大きければ学習データは識別境界から離れていることを意味し，汎化性能が高くなりそうなことが予想できる．よって SVM ではマージンを最大化することを考える．マージンを最大にする重み $\mathbf{w}$ を求める問題は以下のとおりである．

$$\mathbf{w}_{\mathrm{SVM}} = \underset{\mathbf{w}}{\mathrm{argmax}} \left\{ \frac{1}{\|\boldsymbol{w}\|} \min_{i \in \{1, \cdots, n\}} y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) \right\} \tag{3.11}$$

これを解くのは複雑であるため，考え方を変える．はじめに $\mathbf{w}$ の大きさを変更してもマージンの大きさには影響がない．例として定数 α 倍した重み $\alpha\mathbf{w}$ とデータの距離 $|r|_{\alpha\mathbf{w}}$ を計算したものを以下に示す．

$$|r|_{\alpha\mathbf{w}} = \frac{|\alpha w_0 + \alpha \boldsymbol{w}^T \boldsymbol{x}|}{\|\alpha \boldsymbol{w}\|} = \frac{\alpha |w_0 + \boldsymbol{w}^T \boldsymbol{x}|}{\alpha \|\boldsymbol{w}\|} = |r|_{\mathbf{w}} \tag{3.12}$$

そこで識別境界からもっとも近いデータ点を $\boldsymbol{x}_j$ とし

$$y_j (w_0 + \boldsymbol{w}^T \boldsymbol{x}_j) = 1 \tag{3.13}$$

とする．これによって $\mathbf{w}$ の大きさを制限できる．上式を満たすとき

$$y_j (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) \geq 1, \quad \forall i = 1, \cdots, n \tag{3.14}$$

が成り立つ (もっとも近いデータの距離が 1 ならば距離が 1 未満のデータは存在しない)．この条件下で式 (3.10) のマージンは

$$\frac{1}{\|\boldsymbol{w}\|} \min_{i \in \{1, \cdots, n\}} y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) = \frac{1}{\|\boldsymbol{w}\|} \tag{3.15}$$

と非常に簡単になる．$1/\|\boldsymbol{w}\|$ を最大化することは $\|\boldsymbol{w}\|$ を最小化，更には $\|\boldsymbol{w}\|^2$ を最小化することと等価である．このことから式 (3.11) は下式のような制約付き最適化問題に変形できる．

$$\begin{aligned} &\mathbf{w}_{\mathrm{SVM}} = \mathrm{argmin}\, \mathbf{w} \frac{1}{2} \|\boldsymbol{w}\|^2 \\ &s.t. \;\; y_i (w0 + \boldsymbol{w}^T \boldsymbol{x}_i) \geq 1, \quad \forall i = 1, \cdots, n \end{aligned} \tag{3.16}$$

制約付き最適化問題を解くためにはラグランジュの未定乗数法を用いる．各不等式制約に対応する未定乗数を $\boldsymbol{a} = (a_1, \cdots, a_n)^T$ $(a_i \geq 0)$ とおき，ラグランジュ関数を

$$L(\mathbf{w}, \boldsymbol{a}) = \frac{1}{2} \|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i (1 - y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i)) \tag{3.17}$$

と定義する．式 (3.16) は強双対性を満たすことが知られており [2]

$$\mathbf{w}_{\mathrm{SVM}} = \underset{\mathbf{w}}{\operatorname{argmin}} \max_{\boldsymbol{a}|a_i \geq 0} L(\mathbf{w}, \boldsymbol{a}) \tag{3.18}$$

と書き換えられる．これを解くためには最急降下法を用いれば良い．ラグランジュ関数を $\boldsymbol{w}, w_0$, 各未定乗数 a_i で微分すると

$$\begin{aligned}
\frac{\partial L(\mathbf{w}, \boldsymbol{a})}{\partial \boldsymbol{w}} &= \boldsymbol{w} + \sum_{i=1}^{n} a_i (0 - y_i (0 + \boldsymbol{x}_i)) \\
&= \boldsymbol{w} - \sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i \qquad (3.19) \\
\frac{\partial L(\mathbf{w}, \boldsymbol{a})}{\partial w_0} &= 0 + \sum_{i=1}^{n} a_i (0 - y_i (1 + 0)) \\
&= -\sum_{i=1}^{n} a_i y_i \qquad (3.20) \\
\frac{\partial L(\mathbf{w}, \boldsymbol{a})}{\partial a_i} &= 0 + 1(-y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i)) \\
&= 1 - y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) \qquad (3.21)
\end{aligned}$$

となる．初期値 $\mathbf{w}^{(0)}, \boldsymbol{a}^{(0)}$ を設定し上式を用いて最急降下法で更新することで求められる（この時 $\boldsymbol{w}, w_i$ は最小化，a_i は最大化かつ更新は条件 $a_i \geq 0$ を満たすように補整する必要があることに注意する).

ラグランジュの未定乗数に対する KKT 条件 [3] から最適解に対して全ての $i = 1, \cdots, n$ に対して下式が成り立つ.

$$a_i \geq 0 \tag{3.22}$$
$$1 - y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) \leq 0 \tag{3.23}$$
$$a_i (1 - y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i)) = 0 \tag{3.24}$$

式 (3.22),(3.23) は今までに出てきている条件と同じである (未定乗数の制約，最適化問題の制約条件)．式 (3.24) は

$$a_i > 0 \Rightarrow y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) = 1 \tag{3.25}$$
$(1 - y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) = 0$ でなければ等式が成り立たない)
$$y_i (w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) > 1 \Rightarrow a_i = 0 \tag{3.26}$$
($a_i = 0$ でなければ等式が成り立たない)

を意味する．この式からマージンを決定しているのはマージン上のデータ (マージンまでの距離が 1 のデータ) でありこれをサポートベクトルという.

次に，最急降下法の更新式が多く大変なためより簡単にすることを考える．最急降下法で最適解を求めるとき, $\partial L(\mathbf{w}, \boldsymbol{a})/\partial \boldsymbol{w} = 0, \partial L(\mathbf{w}, \boldsymbol{a})/\partial w_0 = 0$ を満たす必要がある．このことと式 (3.19),(3.20) から

$$\begin{aligned}
\boldsymbol{w} - \sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i &= 0 \\
\boldsymbol{w} &= \sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i \qquad (3.27)
\end{aligned}$$

$$-\sum_{i=1}^{n} a_i y_i = 0$$
$$\sum_{i=1}^{n} a_i y_i = 0 \tag{3.28}$$

が成立する．これらをラグランジュ関数に代入すると

$$\begin{aligned}
L(\mathbf{w}, \boldsymbol{a}) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i(1 - y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i)) \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i(1 - y_i w_0 - y_i \boldsymbol{w}^T \boldsymbol{x}_i) \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i - w_0 a_i y_i - \boldsymbol{w}^T a_i y_i \boldsymbol{x}_i \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} w_0 a_i y_i - \sum_{i=1}^{n} \boldsymbol{w}^T a_i y_i \boldsymbol{x}_i \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i - w_0 \sum_{i=1}^{n} a_i y_i - \boldsymbol{w}^T \sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i \quad \text{式 (3.28) を代入} \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i - \boldsymbol{w}^T \sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i \quad \text{式 (3.27) を左向きに代入} \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i - \boldsymbol{w}^T \boldsymbol{w} \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i - \|\boldsymbol{w}\|^2 \\
&= -\frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i \quad \text{式 (3.27) を右向きに代入} \\
&= -\frac{1}{2}\|\sum_{i=1}^{n} a_i y_i x_i\|^2 + \sum_{i=1}^{n} a_i \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j + \sum_{i=1}^{n} a_i = L_{dual}(\boldsymbol{a})
\end{aligned} \tag{3.29}$$

これによってラグランジュ関数が $\boldsymbol{a}$ のみに依存する $L_{dual}(\boldsymbol{a})$ に変換できた．このことから式 (3.16) の最適化問題は $\boldsymbol{a}$ を最適化する以下の問題と等価になる．KKT 条件の制約がある点に注意する．

$$\boldsymbol{a}_{\mathrm{SVM}} = \underset{\boldsymbol{a}}{\mathrm{argmax}} \left\{ \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \right\}$$
$$s.t. \quad \sum_{i=1}^{n} a_i y_i = 0 \text{ かつ } a_i \geq 0, \quad \forall i = 1, \cdots, n \tag{3.30}$$

この最適化問題を解いた後，式 (3.27) から $\boldsymbol{w}$ を求めることができる．式 (3.26) から $y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) > 1$ となる i については $a_i = 0$ となる．よって式 (3.32) の $a_i = 0$ となる学習データは $\boldsymbol{w}$ の決定に寄与せず，$a_i > 0$ となる学習データ（サポートベクトル）のみが $\boldsymbol{w}$ を決定する．ここで，サポートベクトルの添字集合を S と定義する

$$S = \{j | a_j > 0\} \tag{3.31}$$

サポートベクトルはマージンの線上すなわち識別境界から最も近いデータであり，SVM はそれらのみで学習を行うため，外れ値の影響を受けにくいことが分かる．重み $\boldsymbol{w}$ を求める式は以下のように書き換えられる．

$$\boldsymbol{w} = \sum_{i \in S} a_i y_i \boldsymbol{x}_i \tag{3.32}$$

最後に w_0 はどのように求めればよいのか．式 (3.25) から $\forall j \in S$ に対して $y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_j) = 1$ が成り立つ．更に $y_j \in \{-1, 1\}$ なので $y_j^2 = 1$ である．これらから

$$\begin{aligned} y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_j) &= y_j^2 \\ w_0 + \boldsymbol{w}^T \boldsymbol{x}_j &= y_j \\ w_0 &= y_j - \boldsymbol{w}^T \boldsymbol{x}_j \quad \text{式 (3.32) より} \\ &= y_j - \sum_{i \in S} a_i y_i \boldsymbol{x}_i^T \boldsymbol{x}_j \end{aligned} \tag{3.33}$$

誤差を少なくするために平均をとると

$$w_0 = \frac{1}{|S|} \sum_{j \in S} \left\{ y_j - \sum_{i \in S} a_i y_i \boldsymbol{x}_i^T \boldsymbol{x}_j \right\} \tag{3.34}$$

によって w_0 を求められる．

3.2 カーネル関数を用いた非線形な SVM

世の中には線形識別が困難な分類問題が多く存在する．そのような場合には特徴ベクトル $\boldsymbol{x}$ を別の空間に変換することで分離することを考える．関数 ϕ を用いて $\boldsymbol{x}$ を変換したベクトルを $\boldsymbol{z}$ とする．すなわち $\boldsymbol{z} = \phi(\boldsymbol{x})$ である．さらにカーネル関数を $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}') = \boldsymbol{z}^T \boldsymbol{z}'$ と定義する．ここで $\boldsymbol{z}$ に対する線形識別関数

$$g(\boldsymbol{z}) = w_0 + \boldsymbol{w}^T \boldsymbol{z} \tag{3.35}$$

を考えると線形のときの式 (3.29) と同様に $\boldsymbol{z}_i = \phi(\boldsymbol{x}_i)$ とおいてラグランジュの未定乗数法を適用すると

$$\begin{aligned} L_{dual}(\boldsymbol{a}) &= \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \boldsymbol{z}_i^T \boldsymbol{z}_j \\ &= \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \end{aligned} \tag{3.36}$$

を得られる．線形と同様に最急降下法の最適解の条件から

$$\boldsymbol{w} = \sum_{i=1}^{n} a_i y_i \boldsymbol{z}_i \tag{3.37}$$

$$\sum_{i=1}^{n} a_i y_i = 0 \tag{3.38}$$

である．よって最適化問題は下式と等価になる．

$$\begin{aligned} &\boldsymbol{a}_{\mathrm{SVM}} = \underset{\boldsymbol{a}}{\mathrm{argmax}} \left\{ \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \right\} \\ &s.t. \quad \sum_{i=1}^{n} a_i y_i = 0 \text{ かつ } a_i \geq 0, \quad \forall i = 1, \cdots, n \end{aligned} \tag{3.39}$$

重みの計算も線形 SVM と同様に以下のように求められる．

$$\boldsymbol{w} = \sum_{i=1}^{n} a_i y_i \boldsymbol{z}_i \tag{3.40}$$

$$w_0 = \frac{1}{|S|} \sum_{j \in S} \left\{ y_j - \sum_{i \in S} a_i y_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \right\} \tag{3.41}$$

4 ソフト SVM(Soft SVM)

分類したいデータが線形分離不可であったり，カーネル関数が複雑になってしまうときは，どうすればよいのか．SVM ではマージン内への侵入を許容する（分類不可もしくは誤分類を許容するマージンを求める）ことで解決する．これをソフト SVM という．対して今までのものをハード SVM という．ソフト SVM については線形 SVM のときについてのみ説明する．

線形識別関数 $g(\boldsymbol{x}) = w_0 + \boldsymbol{w}^T \boldsymbol{x}$ とする．線形分離が不可なとき線形 SVM における式 (3.14) は成り立たない．そこで新たに $\xi_i \geq 0$ による以下の制約を考える．

$$y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) \geq 1 - \xi_i, \quad \forall i \; 1, \cdots, n \tag{4.1}$$

マージンよりも識別境界に近い位置にあるデータに対して $\xi_i \geq 0$ を与えることで式の辻褄をあわせる．ξ_i に制限がなければどれだけでも入り込める，すなわち $\boldsymbol{w}$ に制約がなくなるのでできるだけ小さい ξ_i を選択するようにしたい．そこでハード SVM の目的関数式 (3.16) を修正し以下のような条件付き最適化問題を考える．

$$\begin{aligned} &\mathbf{w}_{SVM} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{w}\| + C \sum_{i=1}^{n} \xi_i \right\} \\ &s.t. \;\; y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) \geq 1 - \xi_i \text{かつ} \xi_i \geq 0, \;\; \forall i = 1, \cdots, n \end{aligned} \tag{4.2}$$

C は定数であり，大きくすると ξ_i の影響が大きくなり小さくすると影響が小さくなる．すなわち C の大小によって許容する ξ_i の大きさが変わるようになる．

この問題にラグランジュの未定乗数法を適用する．新たな制約 $-\xi_i \leq 0$ に対応する未定乗数を $b_i \geq 0$ とするとラグランジュ関数は

$$L(\mathbf{w}, \boldsymbol{a}, \boldsymbol{\xi b}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} a_i (1 - \xi_i - y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i)) - \sum_{i=1}^{n} b_i \xi_i \tag{4.3}$$

と定義される．このとき $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n), \boldsymbol{b} = (b_1, \cdots, b_n)$ である．ハード SVM と同様に最急降下法の最適解の条件を利用する．はじめにラグランジュ関数を各変数で微分すると

$$\frac{\partial L(\mathbf{w}, \boldsymbol{a}, \boldsymbol{\xi b})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i \tag{4.4}$$

$$\frac{\partial L(\mathbf{w}, \boldsymbol{a}, \boldsymbol{\xi b})}{\partial w_0} = -\sum_{i=1}^{n} a_i y_i \tag{4.5}$$

$$\frac{\partial L(\mathbf{w}, \boldsymbol{a}, \boldsymbol{\xi b})}{\partial \xi_i} = C - a_i - b_i \tag{4.6}$$

となる．それぞれを 0 と置くと

$$\boldsymbol{w} = \sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i \tag{4.7}$$

$$\sum_{i=1}^{n} a_i y_i = 0 \tag{4.8}$$

$$a_i = C - b_i \tag{4.9}$$

が得られる．上記 3 式をラグランジュ関数，式 (4.3) に代入すると

$$\begin{aligned}
L(\mathbf{w}, \boldsymbol{a}, \boldsymbol{\xi b}) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} a_i(1 - \xi_i - y_i(w_0 + \boldsymbol{w}^T\boldsymbol{x}_i)) - \sum_{i=1}^{n} b_i\xi_i \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} a_i\xi_i - w_0\sum_{i=1}^{n} a_i y_i - \boldsymbol{w}^T\sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i - \sum_{i=1}^{n} b_i\xi_i \\
&\quad \text{式 ((4.8)) を代入} \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} a_i\xi_i - \boldsymbol{w}^T\sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i - \sum_{i=1}^{n} b_i\xi_i \\
&\quad \text{式 ((4.9)) を代入} \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} a_i - \sum_{i=1}^{n}(C - b_i)\xi_i - \boldsymbol{w}^T\sum_{i=1}^{n} a_i y_i \boldsymbol{x}_i - \sum_{i=1}^{n} b_i\xi_i \\
&\quad \text{式 ((4.7)) を代入} \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} a_i - \sum_{i=1}^{n}(C - b_i)\xi_i - \boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{n} b_i\xi_i \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} a_i - C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} b_i\xi_i - \|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} b_i\xi_i \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} a_i - \|\boldsymbol{w}\|^2 \\
&= \sum_{i=1}^{n} a_i - \frac{1}{2}\|\boldsymbol{w}\|^2 \\
&= \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j \\
&= L_{dual}(\boldsymbol{a})
\end{aligned} \tag{4.10}$$

が得られる．これはハード SVM と全く同じ式である．しかし式 (4.9) から

$$\begin{aligned}
b_i &= C - a_i \quad b_i \geq 0 \text{ より} \\
C - a_i &\geq 0 \\
a_i &\leq C
\end{aligned} \tag{4.11}$$

という制約がある．これによりソフト SVM は以下の最適化問題になる．

$$\boldsymbol{a}_{\mathrm{SVM}} = \underset{\boldsymbol{a}}{\mathrm{argmax}} \left\{ \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \right\}$$
$$s.t. \quad \sum_{i=1}^{n} a_i y_i = 0 \text{ かつ } 0 \leq a \leq C, \quad \forall i = 1, \cdots, n \tag{4.12}$$

最適解においては KKT 条件から次式が成り立つ [4].

$$a_i(1 - \xi_i - y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i)) = 0 \tag{4.13}$$
$$b_i \xi_i = 0 \tag{4.14}$$

式 (4.13) はハード SVM と同様に以下のことを表す.

$$a_i > 0 \Rightarrow y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) = 1 - \xi_i \tag{4.15}$$
$$y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) > 1 - \xi_i \Rightarrow a_i = 0 \tag{4.16}$$

式 (4.14) は $\xi_i > 0$ すなわち $\boldsymbol{x}_i$ がマージン内にあるとき $b_i = 0$ となることを表しており，このとき式 (4.9) から $a_i = C$ となる．a_i の値によってデータの性質を分けると $a_i = 0$ ならば分類ができており，$0 < a_i < C$ のときそのデータはソフト SVM のサポートベクトルである．そして $a_i = C$ であるデータはマージンよりも内側に存在する.

サポートベクトルの集合を $M = \{i | 0 < a_i < C\}$ と定義する．このときサポートベクトル $i \in M$ に対して $y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) = 1$ が成り立つ．また $y_i^2 = 1$ より

$$y_i(w_0 + \boldsymbol{w}^T \boldsymbol{x}_i) = 1$$
$$y_i(w_0 + \sum_{j=1}^{n} a_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i) = y_i^2$$
$$w_0 = y_i - \sum_{j=1}^{n} a_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i \tag{4.17}$$

となる．また，集合 $S = i | 0 < a$ を定義するとき最適化問題を解いたときの重みの求め方は以下のとおりである.

$$\boldsymbol{w} = \sum_{i \in S} a_i y_i \boldsymbol{x}_i \tag{4.18}$$

$$\boldsymbol{w}_0 = \frac{1}{|M|} \sum_{j \in M} \left\{ y_j - \sum_{i \in S} a_i y_i \boldsymbol{x}_i^T \boldsymbol{x}_j \right\} \tag{4.19}$$

付録 A　ラグランジュの未定乗数法 [1]

最適化したい目的関数

$$f(\boldsymbol{x}) = f(x_1, \cdots, x_d) \tag{A.1}$$

について，K 個の請託条件

$$g_k(\boldsymbol{x}) = 0, \quad k = 1, \cdots, K \tag{A.2}$$

が全て凸関数であるとする．このとき未定乗数 $\lambda_1, \cdots, \lambda_K$ を用意し

$$F(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{k=1}^{K} \lambda_k g_k(\boldsymbol{x}) \tag{A.3}$$

を構成する．$F(\boldsymbol{x})$ の極値問題を解けば，K 個の制約条件のもとで $f(\boldsymbol{x})$ を最大化 (最小化) する $\boldsymbol{x}$ が求まる．

付録 B　最急降下法

極値問題を解くことを考える．関数 $f(\boldsymbol{x})$ の極小値 (極大値) を求めるとき以下の手順で求めることを最急降下法という．

1. 適当な初期値 $\boldsymbol{x}^(0)$ を決める．
2. $\partial f(\boldsymbol{x}^{(k)})/\partial \boldsymbol{x}^{(k)} = 0$ なら終了
3. 下式の通りに $\boldsymbol{x}^{k+1}$を決める(η は定数)

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \eta \frac{\partial f(\boldsymbol{x}^{(k)})}{\boldsymbol{x}^{(k)}} \tag{A.1}$$

4. 2 に戻る

極大値を求める場合，更新式の符号を変えればよい．利点として，凸関数であれば必ず最適解が求まる点，単純なため実装な簡単な点があげられる．欠点としては微分可能な関数でなければ更新できないこと，極値が複数ある場合は局所解に陥るかどうかが初期値依存になることなどがあげられる．

参考文献

[1] 後藤正幸, 小林学, “入門パターン認識と機械学習”, コロナ社, pp134–153,pp227-230,2014

[2] “ラグランジュ関数の背後にある理論 (Boyd 本 5 章概要) - うどん記”(20.10.2016),http://ir5.hatenablog.com/entry/20141214/1418553079

[3] 竹縄知之 (東京海洋大学),“最適化数学-レジュメ 2”,http://www2.kaiyodai.ac.jp/ takenawa/optimization/resume10-2.pdf

[4] 中川裕志 (東京大学),“クラシックな機械学習の入門-サポートベクターマシン”,http://www.r.dl.itc.u-tokyo.ac.jp/ nakagawa/SML1/kernel1.pdf