Implementation of SVM

Taniguchi Taichi

October 5, 2018

Contents

Part I
Introduction

1 目的

今回の最終的な課題を明示してお着ます。

https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+ には８０００個の住宅環境のデータがあり、その住宅に人が住んでいるのかどうかというデータが $(0,1)$ というリストで格納されています。この 8000 個のデータを用いてまず７５００個のデータを学習用に使いその住宅地には人が住んでいるのか判断する学習器関数を作ります。そして残りの５００個のデータを使い、正答率を評価します。

今回は PRML という教材を用いて、SVM について学び、その知識で上記の最終的な課題を達成したいと思います。

References

[1] $https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+$

[2] PRML

Part II
線形 SV 分類　(linear SVM)

2 概要

訓練データが $(x_i, y_i)_{i\in[n]}$ で与えられている問題を扱います。ただし、$[n] = 1, 2, , , n$ とし、$y_i \in (-1, 1)$ とします。SVM での分類器

$g(x)$ は　$f(x) > 0 \implies g(x) = 1$　$f(x) < 0 \implies g(x) = -1$ を取るように定義します。このような関数 $f(x)$ を求めることを考えます。

線形 SV 分類ではこの $f(x)$ を $f(x) = W^T \times x + b$ という一次関数を考えます。

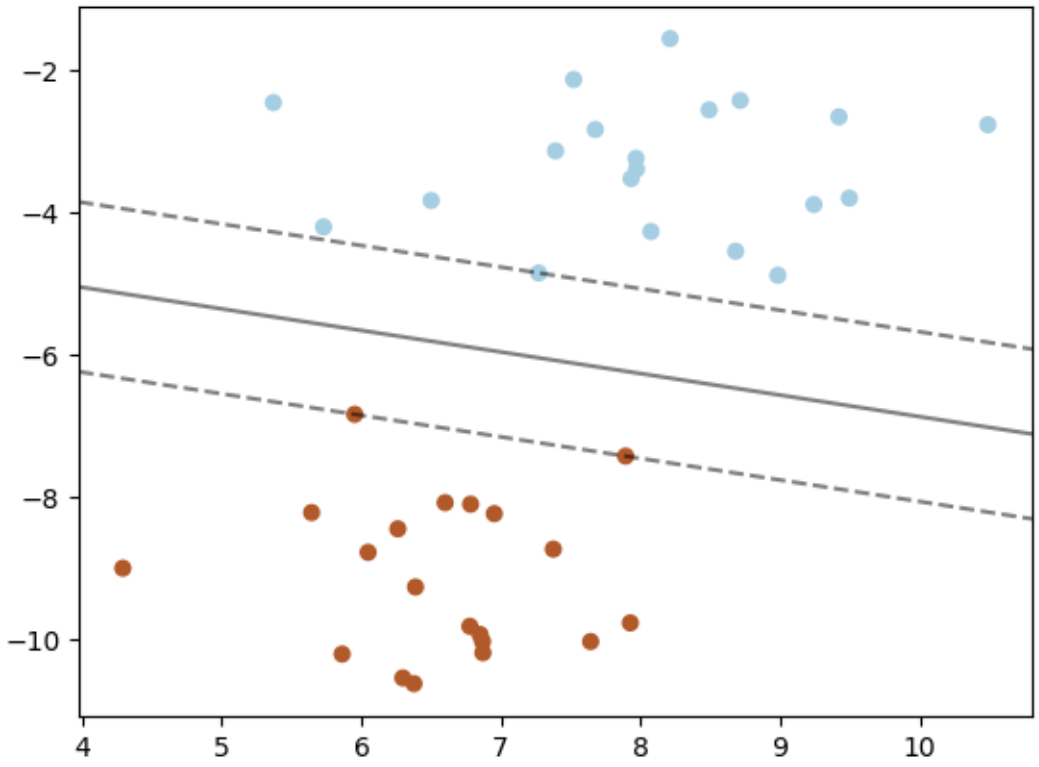
$$f(x) > 0 \implies g(x) = 1$$

$$f(x) < 0 \implies g(x) = -1$$

であるため、分類境界は $f(x) = 0$ となります。W,b は未知のパラメータであり、SVM ではマージンという概念を用いて最適な W,b を求めます。

3 マージン (margin)

分類境界を挟んで計算した、二つのクラス間の距離をマージンといいます。



この画像の場合、二つの点線の距離がマージン (margin) となります。上側の破線がちょうど $f(x) = 1$、下の破線がちょうど $f(x) = -1$ となります。SVM ではこのマージンを最大化させる W,b がもっともよい境界だと定義します。

4 ハードマージン

4.1 ハードマージンの制約条件

ハードマージンでは訓練データを正しく分類できるような W,b が存在すると仮定します。ある y_i について分類が成功しているのであれば y_i と $f(x)$ の符号は一致しているはずです。よって訓練データを完全に $-1, 1$ に分類できる仮定を $y_i f(x_i) > 0$ と表現できます。これを制約条件といいます。

4.2 マージンの最大化

SVM ではマージンを最大化させるような W,b を最適解とするのでした。ではマージンとはどのような式になっているのか考えます。分類境界 $W^t x + b$ とあるデータ点 x_i との距離は次のようにあらわされます。

$$\frac{|W^t x_i + b|}{||W||}$$

そしてこの値を最小にさせる x_i が分類境界と一番近いデータ点となります。

さらにすべてのデータを正しく分類させるという条件から $y_i f(x_i) > 0$ がすべての i で成り立っていなければいけません。ある正の整数 $M > 0$ が $y_i f(x_i) > M$ がすべての i で成り立っているとします。

このときマージンの最大化は次のように表現できます。

$$\max_{w,b,M} \frac{M}{||W||} \quad s.t \quad y_i(W^t x_i + b) \leq M \forall i \in [n]$$

この M は $(W^t x_i + b)$ が最も小さな値と等しくなります。そのようになる i を i_0 とします。

4.3 ハードマージンの目的関数

$$\frac{M}{||W||} = \frac{y_{i_o}(W^t x_{i_0} + b)}{||W||}$$

y_{i_0} は 1 か-1 なので

$$\frac{M}{||W||} = \frac{y_{i_o}(W^t x_{i_0} + b)}{||W||} = \frac{|(W^t x_{i_0} + b)|}{||W||}$$

この式をより簡単にします。W を新たに WM、b を新たに bM と置き換えます。すると

$$\max_{w,b} \frac{M}{||WM||} \quad s.t \quad y_i(W^t M + bM) \geq M \ \ \forall i \in [n]$$

であるので

$$\max_{w,b} \frac{1}{||W||} \quad s.t \quad y_i(W^t + b) \geq 1 \ \ \forall i \in [n]$$

と書き換えることができます。さらに $\frac{1}{||W||}$ の最大化が $||W||$ の最小化と等しいことを考えると

$$\min_{w,b} ||W|| \quad s.t \quad y_i(W^t + b) \geq 1 \ \ \forall i \in [n]$$

これを目的関数として最適化します。

以上がハードマージンの目的関数です。

5 ソフトマージン

5.1 ソフトマージンの制約条件

ハードマージンではすべてのデータが分類境界で、正しく分類できると仮定しました。しかし、実際にはこのような仮定は現実には厳しすぎます。ですのですべてのデータ点ではなく、ある程度分類できたらよしとする条件に変えてみます。それがソフトマージンでの仮定です。

すべてのデータを正しく分類できるという仮定での制約条件は

$$y_i(W^t x_i + b) \geq 1 \ \ \forall i \in [n]$$

でした。つまり、マージン上の図で破線の間にデータ点はなくすべてのデータ点が破線の間にできる領域の外側に分布している状態です。

破線の間の領域にデータ点が存在してもよいとする制約条件は以下のようになります。

$$y_i(W^t x_i + b) \geq 1 - \epsilon_i \forall i \in [n]$$

5.2 ソフトマージンの目的関数（主問題）

破線で囲まれた領域を超えたデータ点を誤分類とします。上の図でゆうと上の青い点が一番下の茶色のデータ点が分布している領域に入ってしまっている状態の点のことで

す。誤分類であるデータについては、$y_i(W^t + b) \leq 0$ を満たしているので $\epsilon_i \geq 1$ を満たしていなければいけません。よって $\sum_i^n \epsilon_i < K \;\; K \in N$ であるならば誤分類の数も K 以下となります。破線で構成された領域に入っているが、誤分類ではないデータ点については $0 \leq \epsilon_i \leq 1$ を満たします。
よって $\sum \epsilon_i$ を小さくすることで誤分類を少なくすることができます。

よって最適化問題は次のように定義しなおされます。

$$\min_{i \in [n]} \frac{1}{2}||W||^2 + C \sum_{i \in [n]} \epsilon_i$$
$$y_i(W^t x_i + b) \geq 1 - \epsilon_i, \forall i \in [n], \epsilon_i \geq 0, \forall i \in [n] \tag{5.1}$$

この問題を SV 分類の主問題と呼びます。(5.1 としておきます。)

C は正則化係数と呼ばれ事前に自分で値を決めておくものです。正則化係数の役割は、$\sum \epsilon$ の抑制度を調節します。C を大きくするとハードマージンに近づきます。$C = \infty$ の時、無限大の値が目的関数に加わるため $\forall i \in [n]$ $\epsilon_i = 0$ 出なければいけません。これは $C = \infty$ の場合においてはソフトマージンとハードマージンは一致するということを示しています。データが分離可能でない場合、目的関数は ∞ の値をとり、最適化の計算ができません。反対に C を小さくしすぎると誤分類をしても目的関数が大きく増えることはないため、より多くの誤分類を許容してしまいます。C の適切な値は交差検証法と呼ばれるものを使い、様々な値を調べてみる必要があります。

6 双対問題

6.1 主変数と双対変数

以下ではより汎用的なハードマージンの定式化を用います。新たな変数 α_i, μ_i を用いて

$$L(w, b, \epsilon, \alpha, \mu) = \frac{1}{2}||W||^2 + C \sum_{i \in [n]} \epsilon_i \quad - \sum_{i \in [n]} \alpha_i (y_i(W^t x_i + b) - 1 + \epsilon_i) - \sum_{i \in [n]} \mu_i \epsilon_i$$

という関数を定義します。これはラグランジュ関数と呼ばれます。w, b, ϵ を主変数と呼び、α, μ を双対変数と呼びます。

6.2 新たな主問題

ラグランジュ関数を双対変数について最大化したものを $P(w, b, \epsilon)$ をします。つまり

$$P(W, b, \epsilon) = \max_{\alpha, \mu} L(W, b, \epsilon, \alpha, \mu)$$

です。ただし、$\alpha_i, \mu_i > 0 \ \ \forall i \in [n]$ です。この関数を主変数について最小化する問題を考えます。

$$\min_{W,b,\epsilon} P(W,b,\epsilon) = \min_{W,b,\epsilon} \ \max_{\alpha,\mu} L(W,b,\epsilon,\alpha,\mu)$$

この問題はソフトマージンで定義した主問題 (5.1) とまったく等価です。

6.3 5.2 との等価性

$$P(W,b,\epsilon) = \frac{1}{2}||W||^2 + C\sum_{i\in[n]} \epsilon_i \quad + \max_{\alpha,\mu}\{-\sum_{i\in[n]} \alpha_i(y_i(W^t x_i + b) - 1 + \epsilon_i) - \sum_{i\in[n]} \mu_i\epsilon_i\}$$

について、もし主変数が制約条件を満たしていないのであれば、$-y_i(W^t x_i + b) - 1 + \epsilon_i > 0, \ -\epsilon_i > 0$ のどちらかを満たしています。よってその i について α, μ のどちらかをどこまでも大きくすることで、ラグランジュ関数は上に有界ではなくなり、最大値が存在しなくなります。反対に、主変数が制約条件を満たしていれば、$-y_i(W^t x_i + b) - 1 + \epsilon_i < 0, \ -\epsilon_i < 0$ を満たしているので、よってこの時 $\alpha > 0, \mu > 0$ より

$$\max_{\alpha,\mu}\{-\sum_{i\in[n]} \alpha_i(y_i(W^t x_i + b) - 1 + \epsilon_i) - \sum_{i\in[n]} \mu_i\epsilon_i\} = 0 + 0 = 0$$

となります。主変数が制約条件を満たしているとき、実行可能といいます。

以上より、実行可能であるとき

$$P(W,b,\epsilon) = \frac{1}{2}||W||^2 + C\sum_{i\in[n]} \epsilon_i$$

となり、元の最適化問題（上記の主問題 (5.1)）が現れます。よって新たに

$$\min_{W,b,\epsilon} P(W,b,\epsilon) \tag{6.1}$$

を主問題とします。(6.1 としておきます。)

6.4 双対問題の定義

次にラグランジュ関数を主変数について最小化したものを $D(\alpha,\mu)$ とします。つまり

$$D(\alpha,\mu) = \min_{W,b,\epsilon} L(W,b,\epsilon,\alpha,\mu)$$

です。この関数を双対変数について最大化する以下の問題を双対問題といいます。

6.5 双対問題の定式

$$\max_{\alpha,\mu} D(\alpha,\mu) = \max_{\alpha,\mu}\{\min_{W,b,\epsilon} L(W,b,\epsilon,\alpha,\mu)\}$$

この双対問題は双対変数のみで表現することができます。
ラグランジュ関数を W, b, ϵ について、偏微分をとります。

$$\frac{\sigma L}{\sigma W} = W - \sum_{i\in[i]} \alpha_i y_i x_i = 0 \tag{6.2}$$

$$\frac{\sigma L}{\sigma b} = -\sum_{i\in[i]} \alpha_i y_i = 0 \tag{6.3}$$

$$\frac{\sigma L}{\sigma \epsilon} = C - \alpha_i - \mu_i = 0 \tag{6.4}$$

L は W の凸二次関数なので極値が最小値となるのはすぐわかります。また、L は b と ϵ に関して一次関数なので、係数が 0 でないなら、L をどこまでも小さくすることができてしまいます。よって係数は 0 でなければなりません。つまり上の (6.3)(6.4) の式を満たしていなければいけません。(6.2)(6.3)(6.4) の指揮をラグランジュ関数に代入していきます。

$$\begin{aligned} L &= \frac{1}{2}||W||^2 + \sum_{i\in[n]} \alpha_i y_i W^t x_i - b\sum_{i\in[n]} \alpha_i y_i + \sum_{i\in[n]} \alpha_i + \sum_{i\in[n]} (C - \alpha_i - \mu_i)\epsilon_i \\ &= -\frac{1}{2}\sum_{i,j\in[n]} \alpha_i\alpha_j y_i y_j x_i^t x_j + \sum_{i\in[n]} \alpha_i \end{aligned}$$

(6.4) について、

$$C - \alpha_i = \mu_i \geq 0$$

より

$$C - \alpha_i \geq 0$$

がわかります。これらをまとめると双対表限は以下のように書けます。

$$\max_{\alpha} -\frac{1}{2}\sum_{i,j\in[n]} \alpha_i\alpha_j y_i y_j x_i^t x_j + \sum_{i\in[n]} \alpha_i$$

$$s.t \quad \sum_{i\in[n]} \alpha_i y_i = 0 \qquad 0 \leq \alpha_i \leq C \quad i \in [n]$$

SVM の双対問題といえばこの問題のことを指すことが多いです。

7 双対性

7.1 弱双対性と強双対性

主問題と双対問題の関係性について確認しておきます。主問題の最適解を W^*, b^*, ϵ^* とし、双対問題の最適解を α^*, μ^* とします。

この時定義より以下の不等式が成り立つことがわかります。

$$\begin{aligned} D(\alpha^*, \mu^*) &= \min_{W,b,\epsilon} L(W, b, \epsilon, \alpha^*, \mu^*) \leq L(W^*, b^*, \epsilon^*, \alpha^*, \mu^*) \\ &\leq \max_{\alpha,\mu} L(W^*, b^*, \epsilon^*, \alpha, \mu) = P(w^*, b^*, \epsilon^*) \end{aligned} \tag{7.1}$$

よって

$$D(\alpha^*, \mu^*) \leq P(w^*, b^*, \epsilon^*)$$

が成り立ちます。

この関係は弱双対性と呼ばれます。

SVM では以下の強双対性と呼ばれる以下の関係があることが知られています。

$$D(\alpha^*, \mu^*) = P(w^*, b^*, \epsilon^*)$$

7.2 鞍点

強双対性より、(7.1) の不等式から以下がわかります。

$$D(\alpha^*, \mu^*) = L(W^*, b^*, \epsilon^*, \alpha^*, \mu^*) = P(w^*, b^*, \epsilon^*) \tag{7.2}$$

また定義から以下の二つの不等式が得られます。

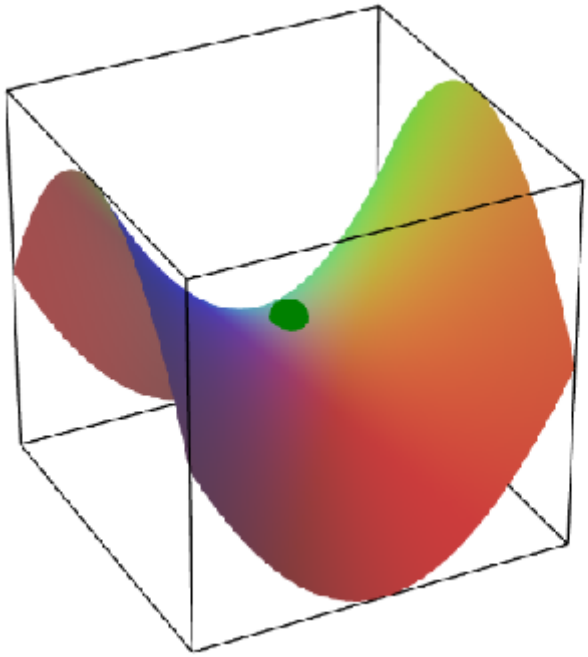
$$P(w^*, b^*, \epsilon^*) = \max_{\alpha,\mu} L(W^*, b^*, \epsilon^*, \alpha, \mu) \geq L(W^*, b^*, \epsilon^*, \alpha, \mu)$$

$$D(\alpha^*, \mu^*) = \min_{W,b,\epsilon} L(W, b, \epsilon, \alpha^*, \mu^*) \leq L(W, b, \epsilon, \alpha^*, \mu^*)$$

この二つの不等式を (7.2) に代入することで以下の不等式を得ます。

$$L(W^*, b^*, \epsilon^*, \alpha, \mu) \leq L(W^*, b^*, \epsilon^*, \alpha^*, \mu^*) \leq L(W, b, \epsilon, \alpha^*, \mu^*)$$

これは $L(W^*, b^*, \epsilon^*, \alpha^*, \mu^*)$ が主変数 w, b, ϵ については極小値で双対変数 α, μ については極大値になることを示しています。このような $(W^*, b^*, \epsilon^*, \alpha^*, \mu^*)$ の点を関数の鞍点といいます。



ある方向では極小値、またある方向では極大値になっています。

8　KKT 条件（Kuhn-Tucker condition）

8.1　KKT 条件とは

KKT 条件とは不等式制約がある関数が最適解をもつための必要十分条件を示します。
関数が最適解をもつためには

$$\frac{\sigma L}{\sigma W} = W - \sum_{i \in [i]} \alpha_i y_i x_i = 0 \tag{8.1}$$

$$\frac{\sigma L}{\sigma b} = -\sum_{i \in [i]} \alpha_i y_i = 0 \tag{8.2}$$

$$\frac{\sigma L}{\sigma \epsilon} = C - \alpha_i - \mu_i = 0 \tag{8.3}$$

$$-\{y_i(W^t x_i + b) - 1 - \epsilon_i\} \leq 0 \tag{8.4}$$

$$-\epsilon_i \geq 0 \tag{8.5}$$

$$\alpha_i \geq 0 \tag{8.6}$$

$$\mu_i \geq 0 \tag{8.7}$$

$$\alpha_i\{y_i(W^t x_i + b) - 1 - \epsilon_i\} = 0 \tag{8.8}$$

$$\mu_i \epsilon_i = 0 \tag{8.9}$$

のすべての条件を満たすことと同値です。した二つの式のことを相補性条件といいます。

8.2 SVM の最適解

KKT 条件から SVM の W, b, ϵ を求めます。双対問題を解いて解 α が求められたとします
W は KKT 条件の一つ目の式から

$$W = \sum_{i \in [i]} \alpha_i y_i x_i$$

であることがすぐにわかります。

次に (8.8) より $\alpha_i > 0$ の時、$y_i(W^t x_i + b) - 1 - \epsilon_i\} = 0$ となります。さらに (8.3) より、$\mu_i = C - \alpha_i$ なので、これを (8.9) に代入すると $(C - \alpha_i)\epsilon_i = 0$ となります。よって $\alpha_i < C$ であるならば、$\epsilon_i = 0$ 出なければなりません。以上をまとめると、

$$\forall i \in [n] \;\; 0 < \alpha_i < C \implies y_i(W^t x_i + b) - 1 = 0$$

この式に W の値を代入して b の値を求めます。

$$
\begin{aligned}
y_i(W^t x_i + b) - 1 = 0 \\
y_i\{(\sum_{k\in[n]} \alpha_k y_k x_k)^t x_i + b\} - 1 = 0 \\
y_i\{\sum_{k\in[i]} \alpha_k y_k x_k^t x_i + b\} - 1 = 0 \\
y_i \sum_{k\in[i]} \alpha_k y_k x_k^t x_i + y_i b = 1 \\
y_i b = 1 - y_i \sum_{k\in[i]} \alpha_k y_k x_k^t x_i \\
b = y_i - \sum_{k\in[i]} \alpha_k y_k x_k^t x_i
\end{aligned}
$$

よって SVM の識別関数は

$$
f(x) = \sum_{i\in[n]} \alpha_i y_i x_i^t x + y_i - \sum_{k\in[i]} \alpha_k y_k x_k^t x_i
$$

となります。

References

[1] $http://scikit-learn.org/stable/_images/sphx_glr_plot_separating_hyperplane_001.png$

[2] $http://mathinsight.org/media/applet/image/large/saddle_point_two_variables.png$