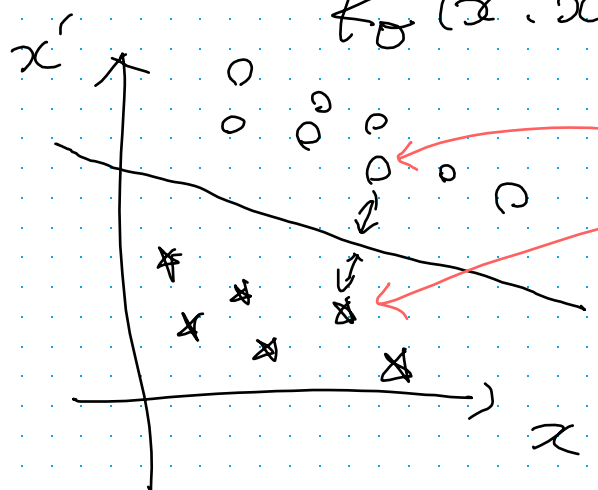


kernel band width σ
 How to chose σ this param

for SVM Classifier

$$k_{\sigma}(x, x') := \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right)$$



$$f(x, x') = w_1 x + w_2 x' + b \\ = w x + b$$

$$\rho = \frac{2}{\|w\|} = \frac{2}{\sqrt{w_1^2 + w_2^2}}$$

この最大化

$\frac{1}{2}\|w\|$ の最小化

非線形にすための $x \mapsto \phi(x)$

$$f(x) = \sum_{i=1}^n \alpha_i y^{(i)} x x' + b \mapsto \sum_{i=1}^n \alpha_i y^{(i)} \underbrace{\phi(x) \phi(x')}_{k_{\sigma}(x, x')} + b$$

カーネル

$$k_0(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

識別関数 $f(x) = xw + b$

$$= \sum_{i=1}^n \alpha_i y_i \underbrace{\phi(x) \phi(x')}_{k_0(x, x')} + b$$

argmin $\int \frac{\|w\|^2}{2}$ Subject to

$$\underbrace{y_i}_{\uparrow} (b + \underbrace{w x_i}_{\uparrow}) \geq 1$$

positive negative

$$z - |x| = 1$$

$$\text{or } |x| = 1$$

カーネル = 内積カーネル

$$f(x) = \sum_{n=1}^N \alpha_n K(x, x_n)$$

無限次元空間で、 N 個の点で

空間の点を完全に区別できる

例えば、

 $x = x_0$ の周りに Training sample $x_1, x_2, \dots$

が存在

と、 x_0 との距離が一番近いか $\phi_{x_1}(x_0), \phi_{x_2}(x_0), \dots, \phi_{x_N}(x_0)$ の中で

一番大きいものを 見つけることができるか?

??

$$K(x, x') = \exp\left(-\frac{|x - x'|_2^2}{2\sigma^2}\right) \quad \sigma = -\frac{1}{2\sigma^2}$$

$$= \exp(-r |x - x'|^2)$$

分散が大きい、 σ が大きい

原因が、分散を小さく

 σ が小さい $\Rightarrow \sigma$ が大きい $\Rightarrow$ 決定境界がざっくり

$$f(x) = e^x \quad \dots \quad \text{e.g.}$$

$$= \sum_{n=0}^{\infty} \left(\frac{f^{(n)}(0)}{n!} \right) \times x^n$$

$$= \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

$$\exp \left(\frac{\|x - x'\|^2}{2\sigma^2} \right)$$

↑
 $\rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow \rightarrow$

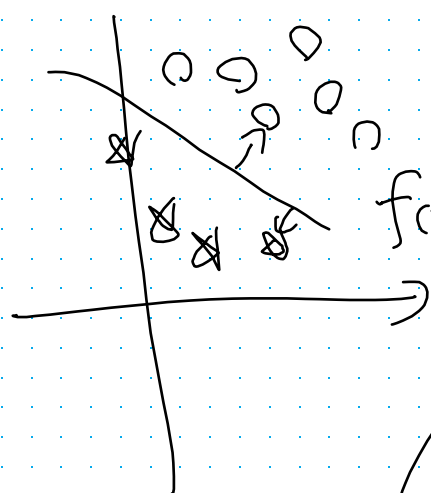
$$\text{fz} \quad \exp \left(\frac{x^T x'}{\sigma^2} \right) = \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{x x^T}{\sigma^2} \right)^n$$

$$= \exp \left(\frac{-x x^T}{2\sigma^2} \right) \exp \left(\frac{x^T x'}{\sigma^2} \right) \exp \left(\frac{x'^T x'}{2\sigma^2} \right)$$

無限のオーダー - カ" $\exp \left(\frac{\|x - x'\|^2}{2\sigma^2} \right)$

つまり $f_0(x, x')$ は 有限のオーダー
 の計算になる

Let us assume hard-margin SVM
 we would like to address the problem of
 maximize margin



$$f(x) = w^T x + b = 0$$

$$\frac{2}{\|w\|}$$

subject to

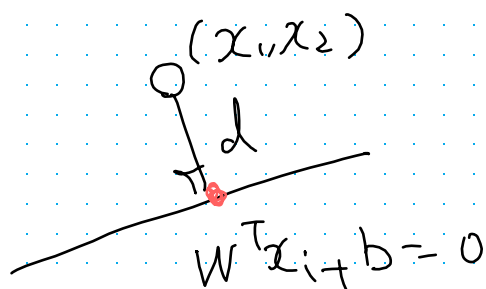
$$w^T x_i + b > 1 \quad x_i \in \text{Class-1}$$

$$w^T x_i + b < -1 \quad x_i \in \text{Class-2}$$



$$\rightarrow x_i (w^T x_i + b) > 1$$

点と直線の距離の公式)



$$d = \frac{|w_1 x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}}$$

Appendix

$A(x_1, y_1)$

$$ax + by + c = 0$$

$$y = -\frac{a}{b}x - \frac{c}{b} \quad \leftarrow \begin{matrix} \text{ㄋㄟ} \\ \text{ㄋㄟ} \end{matrix}$$

$(-\frac{b}{a})$ に平行な直線

AH は $ax + by + c = 0$ の法線ベクトルに等しい

$(\frac{a}{b})$ に平行な直線は

“直交する”

$$AH: \begin{cases} x: x_1 + at \\ y: y_1 + bt \end{cases}$$

媒介変数表示

$$H(x_1 + ah, y_1 + bh) \text{ とおく}$$

H は直線上の点なので

$$a(x_1 + ah) + b(y_1 + bh) + c = 0$$

$$\Leftrightarrow h = \frac{-c - ax_1 - by_1}{a^2 + b^2}$$

$$d^2 = (ah)^2 + (bh)^2 = h^2(a^2 + b^2)$$

ㄋㄟ

$$d = \sqrt{a^2 + b^2} \cdot |h|$$

$$= \frac{|ax_1 + by_1 + c|}{\sqrt{a^2 + b^2}}$$

then maximize $\frac{2}{\|w\|^2}$

can be transformed minimize $\frac{1}{2} \|w\|^2$

subject to $y_i(w x_i + b) > 1$

$\arg \min_w \left\{ \frac{1}{2} \|w\|^2 \right\}$ optimize

by using Lagrange multiplier method

$$\lambda = (\lambda_1, \dots, \lambda_n)$$

$$\lambda_i \geq 0 \quad (i = 1, \dots, n)$$

$$L_p(w, b, \lambda) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \lambda_i (y_i (w x_i + b) - 1)$$

$\Downarrow$

$\arg \min_w \left\{ L_p \right\}$ subject to $\lambda_i \geq 0$

after considering

$$\frac{\partial L_p}{\partial w}, \frac{\partial L_p}{\partial b}, \frac{\partial L_p}{\partial \lambda} \quad , \quad L_p(\lambda) = \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j$$

so $\arg\max_{\lambda} \{LP\}$

by using grid search

$$w = \sum_{i=1}^n \lambda_i y_i x_i$$

or sequential minimal
optimization
(SMO)

in this phase

if we use training sample itself.

to maximize margin

σ become small

which means overfitting

it's a rookie mistake!

↑
typical

はてな P14

①

線形判別関数のマージン最大化

$$\left(\underset{\uparrow}{w^T x_i} + \underset{\nearrow}{b} \right) \geq \kappa$$

 κ を正規化したものに $\kappa=1$ を与える

$$t_i (w^T x_i + b) \geq 1 \quad \dots \star$$

クラス間マージンは $\frac{2}{\|w\|}$ で与えられる最近の超平面の式 $w_0^T x + b_0 = 0$ とすると 最大化したい

この超平面は最大クラス間マージンとなる

最適判別超平面は 2つのクラスの中心

$$\{w_0, b_0\} = \arg \min_{w, b} \frac{1}{2} \|w\|^2$$

不等式制約条件最適化問題

→ 後の章で

の主要問題を解く

この方程式

→ ラグランジアン法

$$L_p(w, b, \lambda) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \lambda_i (t_i (w^T x_i + b) - 1)$$

$$\lambda_i \geq 0$$

$$\lambda_i (t_i (w^T x_i + b) - 1) = 0 \quad \left. \vphantom{\lambda_i} \right\} \text{K}$$

マージンを最大にするための最適解は $t_i (w^T x_i + b) = 1$

それ以外の場合は $\lambda_i = 0$

$$w_0 = \sum_{i=1}^N \lambda_i t_i x_i \quad \text{これは 7.1.7 条件の偏微分から}$$

$\Downarrow$ これは LP に代る

$$L_d(\lambda) = \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j t_i t_j x_i^T x_j$$

dual

最適問題

最適解 λ は $L_d(\lambda)$ を最大にする λ により得られる

$\lambda_i > 0$ とおき x_i をサポートベクトルとみる

最適識別超平面を構成する乗数とみる

最適バイアス b_0 は、サポートベクトル x_s を用いて

$$\omega_0^T x_s + b_0 - 1 = 0 \quad \text{を解く} \quad \wedge$$

Σ の平均を用いる。

$$\max_{\lambda} \gamma = \frac{1}{\|\omega_0\|} = \frac{\gamma}{\sqrt{\omega_0^T \omega_0}}$$

$\bar{\lambda}_i$ は最適解とみる

$$(1 - t_i b_0) = (t_i x_i \omega_0)$$

$$\omega_0^T \omega_0 = \sum_{i=1}^N \bar{\lambda}_i t_i x_i \omega_0$$

$$= \sum_{i=1}^N \bar{\lambda}_i - b_0 \sum_{i=1}^N \bar{\lambda}_i t_i$$

$$= \sum_{i=1}^N \bar{\lambda}_i$$

$$\textcircled{1} \lambda_i > 0 \quad \text{if} \quad t_i(\omega_0 x_i + b) = 1$$

$$\textcircled{2} \lambda_i = 0 \quad \text{if} \quad t_i(\omega_0 x_i + b) > 1$$

$$\Rightarrow \frac{1}{\|\omega_0\|} = \frac{1}{\sqrt{\sum_{i=1}^N \bar{\lambda}_i}}$$

$$\left(\sum_{i=1}^N \hat{\lambda}_i - b_0 \underbrace{\sum_{i=1}^N \hat{\lambda}_i x_i}_{=0} \right) = w_0^T w_0$$

$\therefore$ KKT condition x_i

$$\frac{\partial L_P}{\partial b} = \sum_{i=1}^N \hat{\lambda}_i x_i = 0 \quad \text{if}$$

SVMは 解か学習データの線形結合で表せる
識別超平面

$x \mapsto \phi(x)$ に非線形写像する

d 次元の学習データ $x \in \mathbb{R}^d$ と

その非線形写像の集合 $\{\phi_j(x)\}_{j=1}^M$ を考える

$$\phi(x) = (\phi_0(x), \phi_1(x), \dots, \phi_M(x))^T$$

where $\phi_0(x) = 1$ (バイアス項とる)

非線形特徴空間での線形識別関数

$$h(\phi(x)) = \sum_{j=0}^M \omega_j \phi_j(x) = \omega^T \phi(x) \quad \text{で表す}$$

このとき非線形空間内で SVM を考えると 最適識別超平面は

$$w_0 = \sum_{i=1}^N \lambda_i x_i \phi(x_i) \quad \text{と表す.}$$

このとき識別関数は

$$h(\phi(x)) = w^T \phi(x) = \sum_{i=1}^N \lambda_i x_i \underbrace{\phi(x_i)^T \phi(x)}_{\text{核関数}}$$

核関数: ϕ -内積関数: $k(x, x)$

(ハード) ソフトマシンの識別器の
ラグランジュ未定乗数 λ_i は

次の問題を解くことにより得られる。

対称

$$\max_{\lambda} L(\lambda) = \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j x_i x_j \underbrace{\phi(x_i) \phi(x_j)}_{k(x_i, x_j)}$$

subject to $\lambda^T \mathbf{1} = 0$

↑

ラグラン行列: $N \times N$ の対称行列

次のように 非線形空間の次元が M である場合

$\phi(x_i) \phi(x_j)$ は x_i と x_j の d 次元空間での
内積内積で置き換える

e.g. Radius Based Function Kernel (RBF)

$$k_0(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right)$$

個々のサポートベクターが独立に識別境界を決定するようになる

σ はカーネル関数の広がり（半径）を制御するパラメータ

識別関数

$$h(\phi(x)) = \sum_{i=1}^N \lambda_i \star_i \underline{\kappa(x_i, x)}$$

$$0 \leq \lambda_i \leq \frac{1}{N}$$

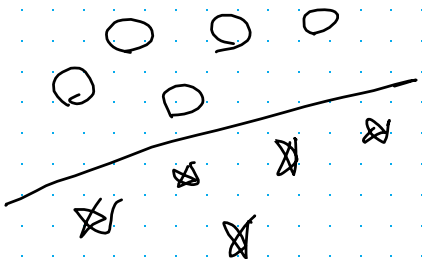
この区間で
なり複雑な
識別曲面が
構成できる

overfitting

では

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

σ が大きい場合はサポートベクター x が遠く離れている
広範囲のサポートベクターが識別に寄与する



σ が大きい



σ が小さい

σ が小さい
⇒ 訓練誤差 ①
⇒ 汎化誤差 ②

$$\begin{aligned} \kappa(x, x') &= \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|x'\|^2}{2\sigma^2}\right) \exp\left(\frac{2x \cdot x'}{2\sigma^2}\right) \\ &= \underbrace{g(x) g(x')}_{\text{指数関数}} \exp\left(\frac{\kappa_1(x, x')}{\sigma^2}\right) \end{aligned}$$

指数関数

$$\kappa(x, x') = \prod_{i=1}^d \exp(u_i, u'_i)$$

無限の次元 $n \rightarrow \infty$