+

サポートベクトルマシーンについて

Kumamoto University

熊本大学
井村　祥太朗

SVM（サポートベクターマシーン）

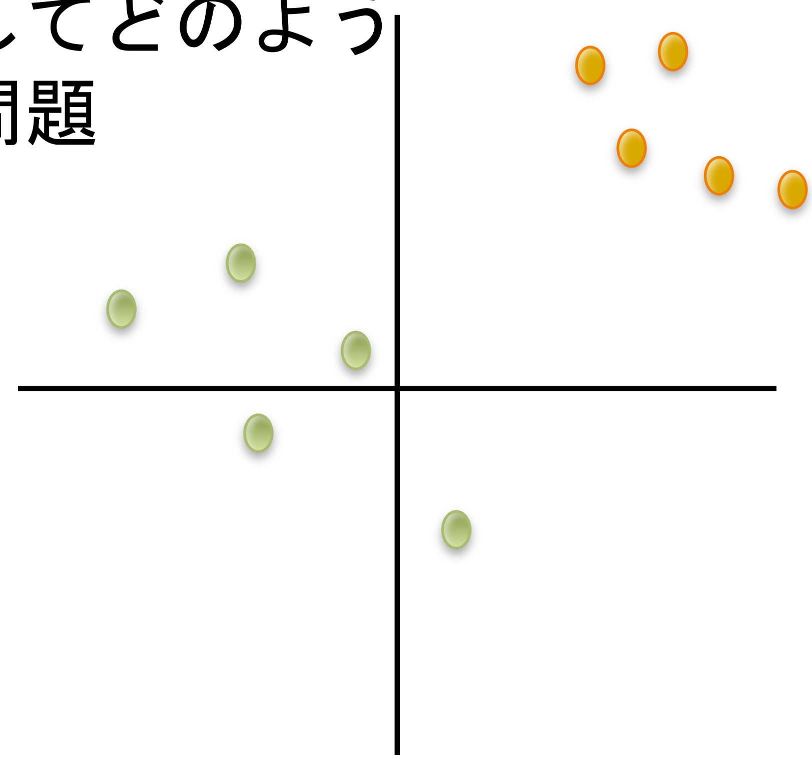
- 1995年にV.Vapnikによって統計的学習理論の枠組みで提案された機械学習

- 学習データの中の，サポートベクトル (識別境界近傍に位置する学習データ) と識別境界との距離であるマージンを最大化するように識別境界を構築し２クラス分類を行う

2クラス分類問題

SVMはパターン認識手法として用いられる
他のアルゴリズムと比べてどう違うのか？

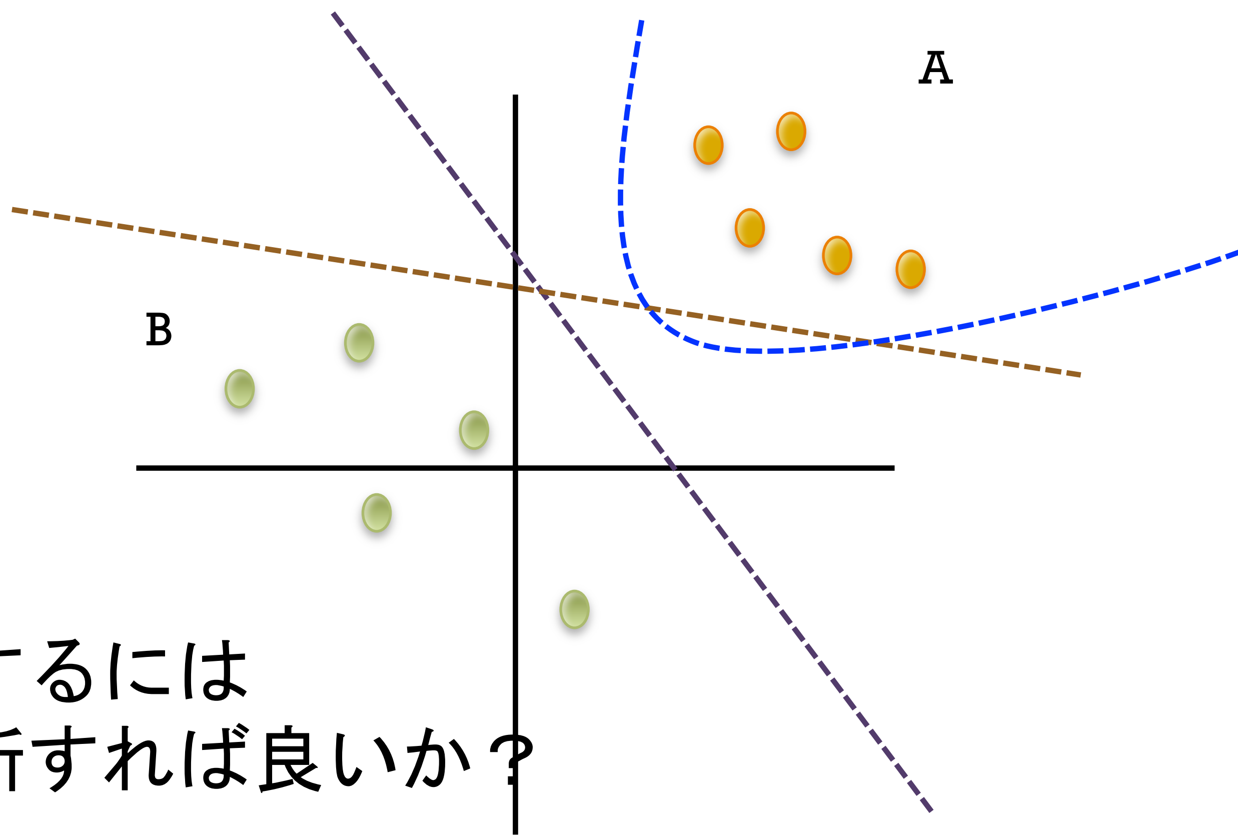
右の図のような２次元空間があった際に
２つのグループを識別するとしてどのよう
識別線を引くと良いかという問題



2クラス分類

識別線として考えられるものは無数に存在する

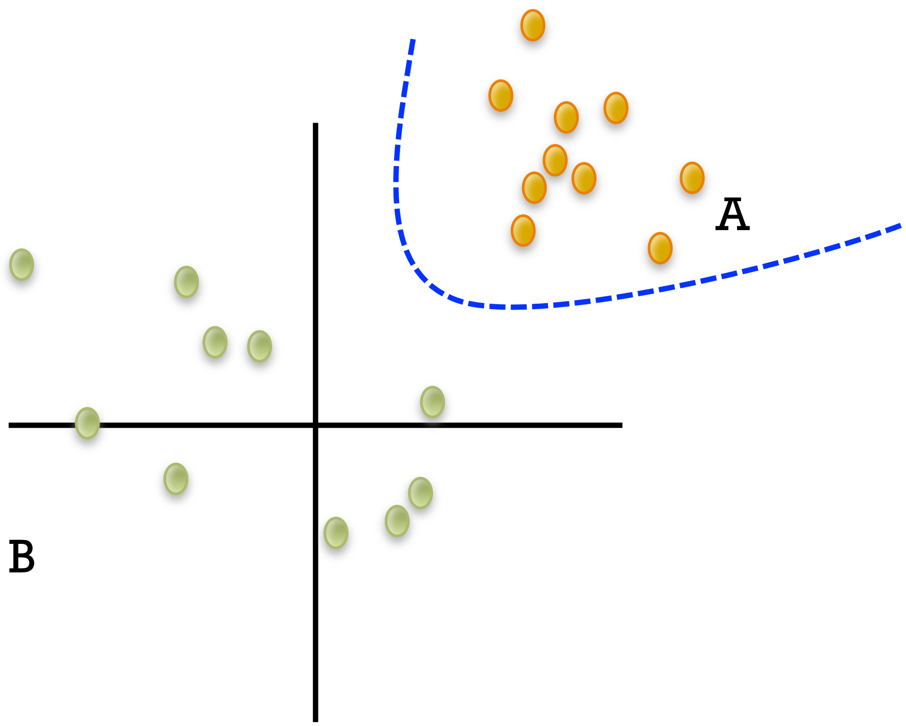
例えば…



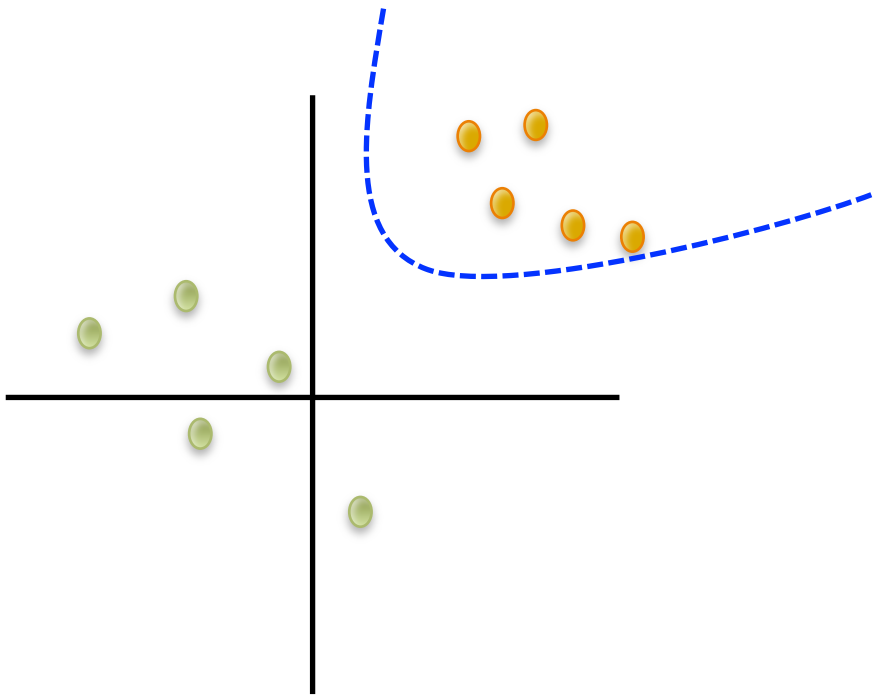
パターン認識するには
何を持って判断すれば良いか？

2クラス分類

バックプロパゲーション学習（従来手法？）だと…



教師データ

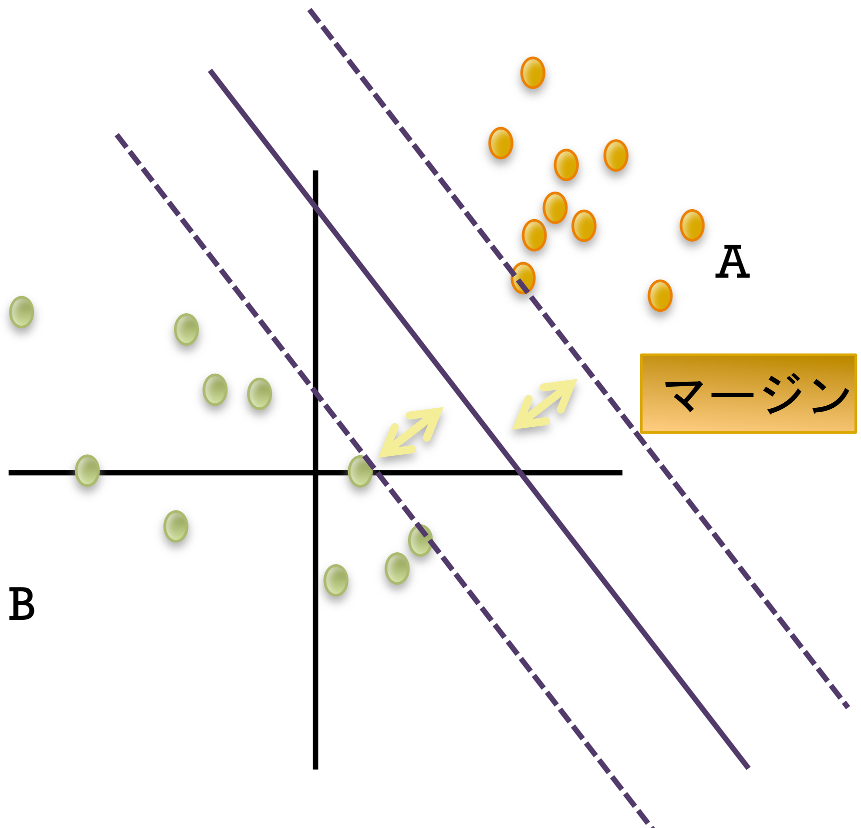


識別データ

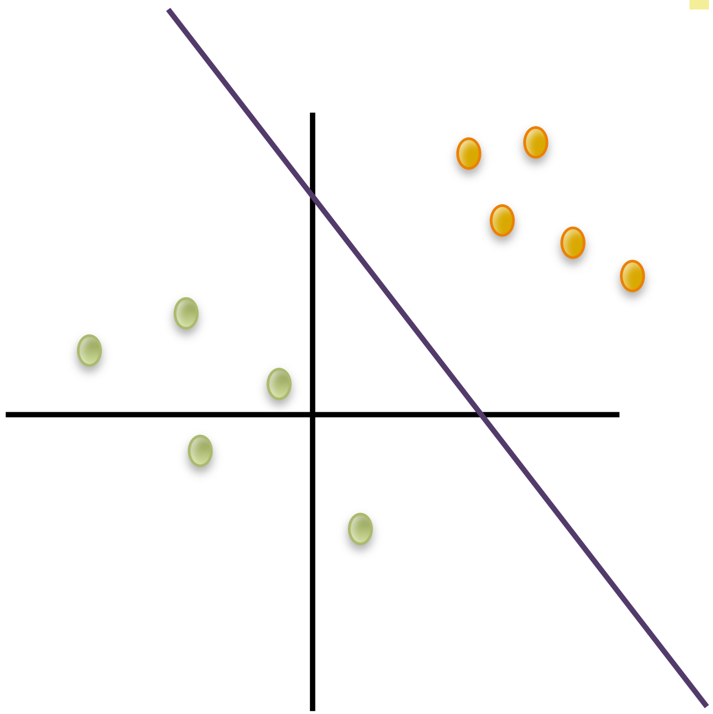
学習した教師データに依存した結果になる
（パラメトリック）

2クラス分類

一方SVMを用いた場合



教師データ



識別データ

学習した教師データから他クラスに一番近いものを基準
（ノンパラメトリック）

超平面集合

- SVMの中核：超平面識別関数
- 内積空間F及びパターンベクトル集合

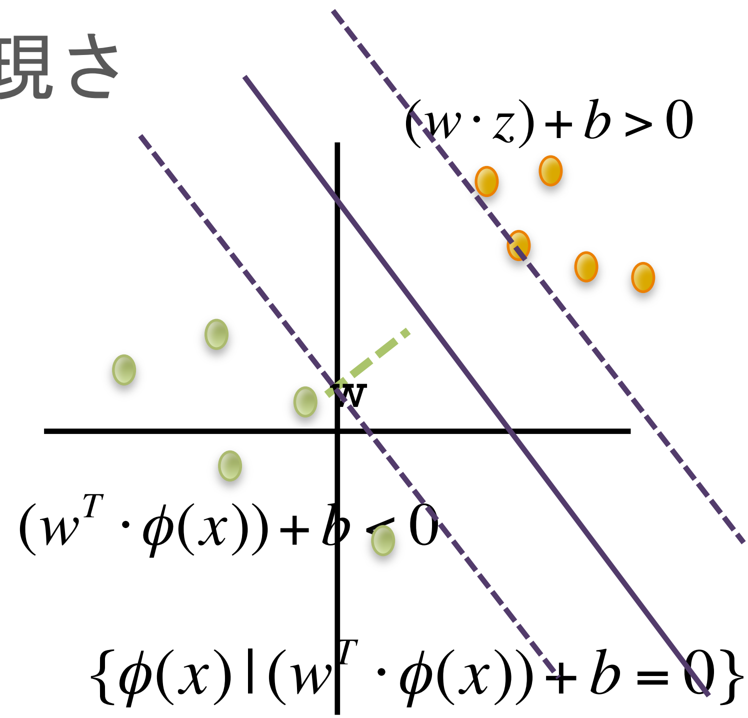
$x_1, \ldots, x_r$ が与えられたとすると任意の超平面識別関数は次のように表現される

$$\{x \in R^P : (w^T \cdot x) + b = 0\}$$

R^P 内積空間

w^T ベクトルパラメータ

b バイアスパラメータ



超平面集合

- 制約を加えることで識別関数となる

超平面を$(w^T, b) \in R^R \times r$ を有する関数

に一意に決める

$$\min_{i=1,\cdots,r} |(w \cdot z_i) + b| = 1$$

この制約によって $w^T b$ は距離

$1/\|w^T\|$ を持つ超平面に接近

するデータを表す

従って２クラス分類問題の場合

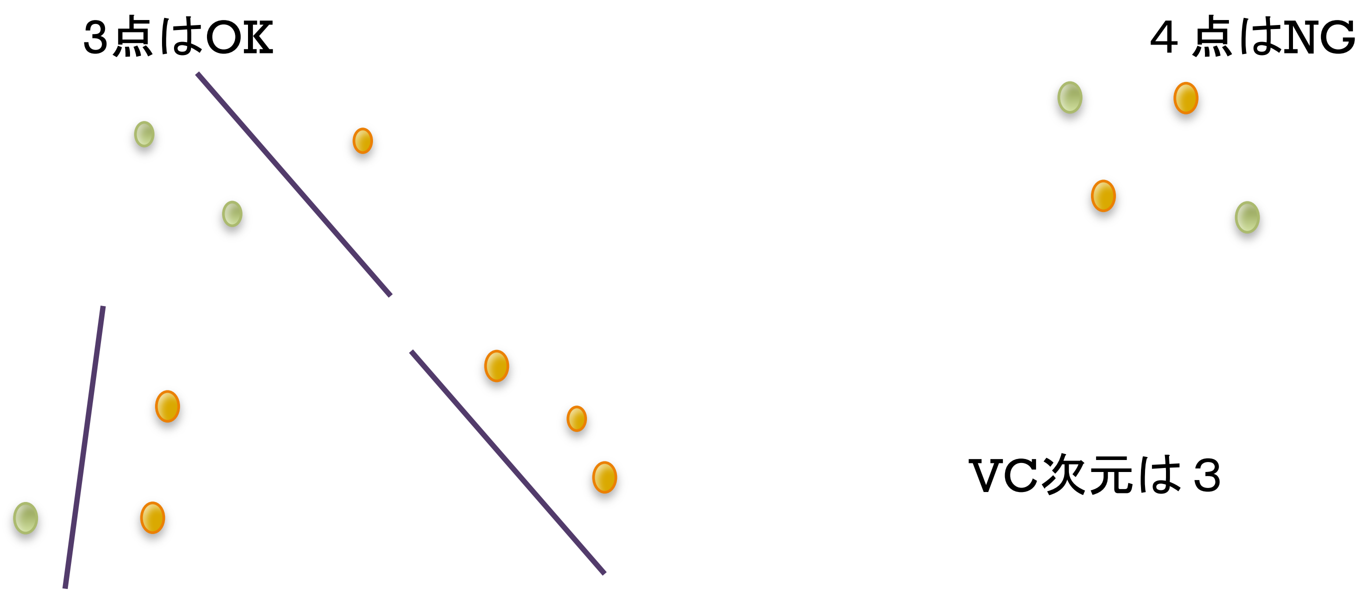
垂直に測ったマージンは少なくとも

$2/\|w\|$ となる

VC次元

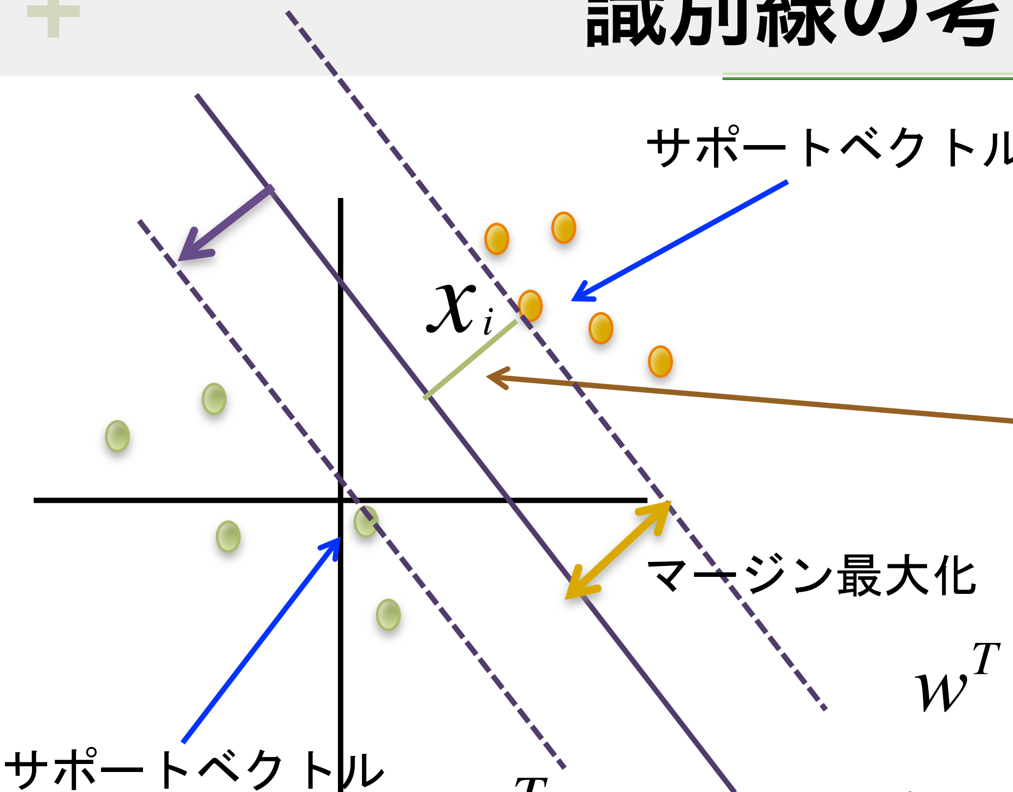
- VC(Vapnik-Chervonenkis)次元：任意の２値ラベルを付与しても正しく分離できるデータ点の最大数
- N次元のときVC次元はN+１

２次元空間内出の線形識別例



VC次元は３

識別線の考え方



$$y = w^T \cdot \phi(x_i) + b$$

超平面への距離

$$d_i = -\frac{w^T \cdot \phi(x_i) + b}{\| w^T \|}$$

$$w^T \cdot \left(\phi(x_i) + d_i \frac{w^T}{\| w^T \|}\right) + b = 0$$

垂線の足

サポートベクトルと呼ばれるクラス境界近傍に位置する訓練点を基準として、その距離が最も大きくなるような位置に識別境界を設定。このことをマージン最大化という。

マージン最大化による推定

入力 $x \in R^p$

出力 $y \in \{+1, -1\}$

- 2値ラベルを±1と扱うことで学習アルゴリズムの記述が楽になる

写像 $\phi : R^p \to H$ 高次元空間 H における関数 $w^T \phi(x) + b$ の値の正負データで分割

$$y_i = +1 \Rightarrow w^T \phi(x) + b > 0$$

$$y_i = +1 \Rightarrow w^T \phi(x) + b < 0$$

マージン最大化による推定

分けられない時でも高次元空間Hなら分けられる可能性がある

ある超平面　$w^T\phi(x)+b=0$ が与えられたとき、超平面に関する学習マージンとは学習データから超平面までの最短距離で定義される

データ $\phi(x_i)$ から超平面 $w^T\phi(x)+b=0$ までの距離は $\dfrac{w^T\phi(x_i)+b}{\|w\|}$

よって超平面までのマージンは $\displaystyle\min_{i=1,\ldots,n}\frac{w^T\phi(x_i)+b}{\|w\|}$ に等しい

判別関数の推定手順

- マージン最大化による判別関数の推定手順は

1. 学習データが線形分離可能という条件下で、**マージンを最大化するような超平面**を求める
2. テスト点が超平面のどちら側にあるかによってラベルの予測を行う

判別関数の推定手順

(x_i, y_i) が $w^T\phi(x_i)+b=0$ によって正しく判別されるとき

$$y_i=+1 \Rightarrow w^T\phi(x_i)+b>0$$

$$y_i=+1 \Rightarrow w^T\phi(x_i)+b<0$$ が満たされる。

このとき $y_i(w^T\phi(x_i)+b)>0$ …① が満たされている。

よって、全てのパラメータに対して①が成り立つことで

線形分離可能であるための必要十分条件となる。

このことからすべての学習データについて①を満たすような

超平面の中からマージンを最大化するものを求めればよい。

最適化問題

最適化問題として書くと

$$\max_{w,b} \min_{i=1,\ldots,n} \frac{w^T\phi(x_i)+b}{\|w\|} \quad subject \quad to \quad y_i(w^T\phi(x_i)+b)>0$$

$$i=1,\ldots,n$$

解を $\hat{w},\hat{b}$ とすると判別器 $f: R^P \to \{+1,-1\}$ は

$$f(x)=\mathrm{sgn}(\hat{w}^T\phi(x)+\hat{b})$$

ここでsgn(z)は

$$\mathrm{sgn}(z)=\begin{cases} +1, z\geq 0 \\ -1, z\leq 0 \end{cases}$$

で定義される符号関数である。

最適化問題

超平面パラメータ w, b をともに定数倍しても

$w^T \phi(x) + b = 0$ は不変であることに着目すると

$$\min_{i=1,\ldots,n} | w^T \phi(x_i) + b | = 1$$

とう条件を加えても、判別関数の推定に影響を与えない

このとき $y_i(w^T \phi(x_i) + b) \geq 1$ が成り立つと最適化問題は

$$\min_{w,b} \frac{1}{2} \| w^2 \| \quad subject \quad to \quad y_i(w^T \phi(x_i) + b) \geq 1 \qquad i = 1, \ldots, n$$

$$\min_{i=1,\ldots,n} | w^T \phi(x_i) + b | = 1$$

に帰着する

最適化問題

ここで　$\max_{w,b} \frac{1}{\|w\|}$ を解きやすいよう$\min_{w,b} \frac{1}{2}\|w\|^2$とする

$$\min_{i=1,\ldots,n} | w^T\phi(x_i) + b = 0 | = 1$$

という制約式を除いても最適解は変わらない

したがって解くべき問題は

$\min_{w,b} \frac{1}{2} \| w^2 \| \quad subject \quad to \quad y_i(w^T\phi(x_i)+b) \geq 1 \quad i = 1,\ldots,n$ となる

するとこの式は線形制約のもとで凸の２次関数最適化問題になり、

有効制約法や内点法などによって最適解を効率的に求めることができる

ハードマージン

線形制約下での凸な２次関数の最適化

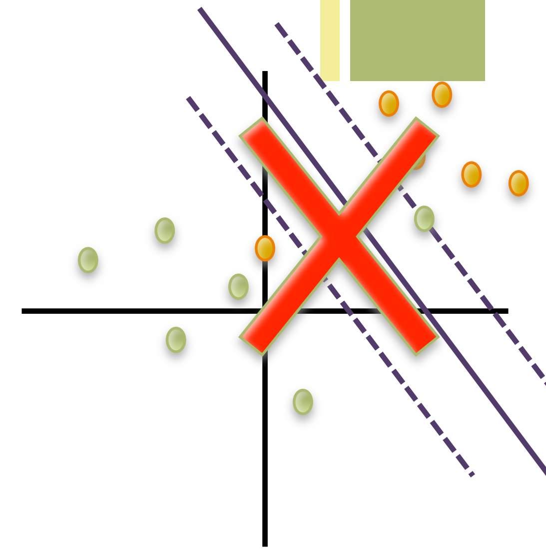
（制約条件を満たさないものを許さない）

- ローカルミニマムの問題がない

（局所最適＝大域的最適）

- 計算パッケージなどで比較的容易に解ける

- このままでも解くことはできるが，双対問題に変換すると，制約の部分がより簡単な式で表される



ソフトマージン

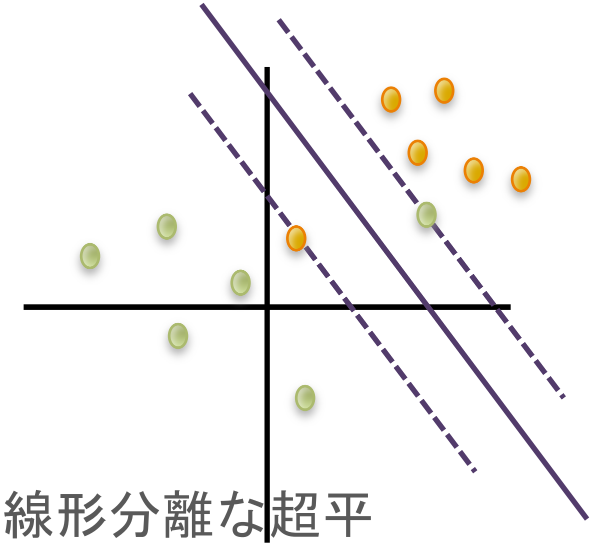
- 訓練データが線形分離可能でないとき,

HSVMは解を持たない少しの誤差 ξ_i を許す

- ペナルティを与えることで, データを

上手く分けられない場合でも“ほぼ”

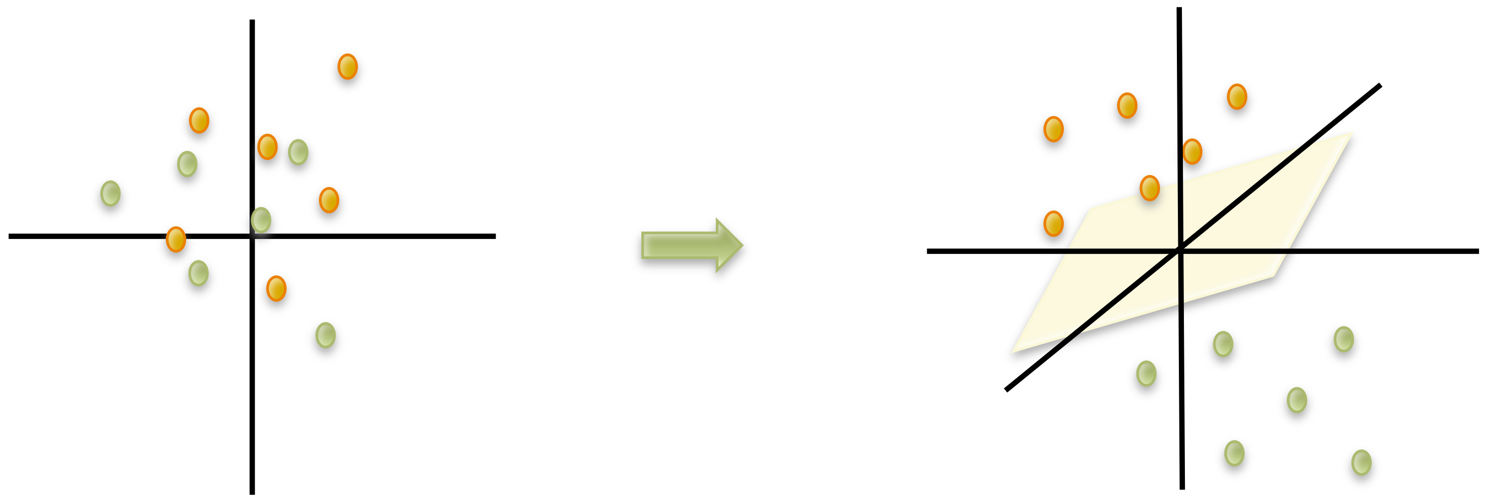
分けられるように超平面を定める.



- 線形分離な場合でもノイズがある場合などは無理に線形分離な超平面を求めないほうがよい

線形識別できない場合

線形で分離できない場合...



高次元特徴空間へ写像（カーネルトリック）

特徴空間で線形SVMを行う

SVMを用いるメリット・デメリット

- メリット
- データ特徴量の次元が増加しても識別精度がよく,最適化するパラメータが少ない
- 最適化すべき パラメータが少ないので試行回数が少なく最適パラメータを求められる

デメリット

- 学習データが増えると計算量が膨大
- 基本的には 2クラスの識別手法で 最適パラメータが異なり,多クラスを考慮に入れた識別関数の最適化・超平面作成は困難

参考文献

- 金森敬文,竹之内高志,村田昇

2009　共立出版　「Rで学ぶデータサイエンス 5 パターン認識」

- 杉山 将　東京工業大学

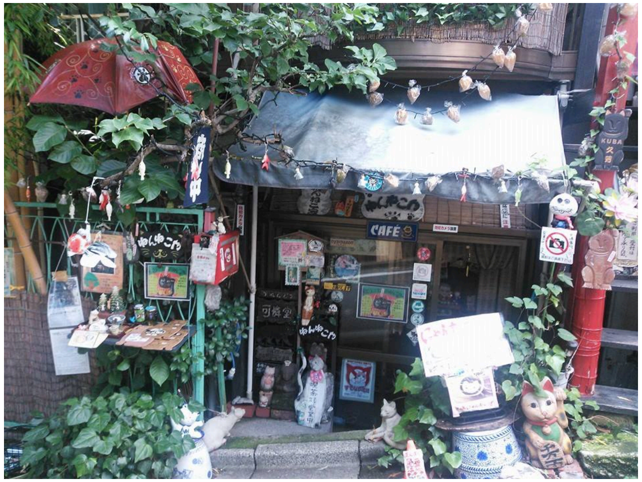
2011　「サポートベクターマシンとブースティングの学習理論」

- 赤穂昭太郎　産業技術総合研究所

2006　統数研公開講座「カーネル法の最前線—ＳＶＭ, 非線形データ解析,構造化データ—」

- 小野田崇

2001　オペレーションリサーチ　「サポートベクターマシーンの概要」



ご清聴と12日間

お世話になりました

ありがとうございます

カーネル化SVM

最適化

$$\max_{\alpha_1,\ldots,\alpha_n} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j k(x_i \cdot x_j)$$

$$subject \quad to \quad 0 \le \alpha_i \le C, \sum_i y_i a_i = 0$$

サポートベクトル

$$S : 0 \le \alpha_i \le C \qquad O : \alpha_i = C \qquad I : \alpha_i = 0$$

識別関数

$$y = \mathrm{sgn}[\sum_{i \in S \cup O} a_i k(x_i, x) + b]$$

閾値

$$b = y_i - \sum_j a_i k(x_j, x_i) \qquad (i \in S)$$

補足

$$\min_{i=1,\ldots,n} \frac{w^T \phi(x_i) + b}{\| w \|}$$

$$\frac{1}{\| w \|} \rightarrow \max$$

$$\| w^2 \| \rightarrow \min$$