第7章　疎な解を持つカーネルマシン

Pattern Recognition and Machine Learning

修士２年
山川佳洋

BehaviorInNetworks
都市生活学・ネットワーク行動学研究室

本章の概要

- SVM (Support vector machine)

－識別関数の一種(出力の事後確率は不明)

－モデルパラメータの局所解が大域解になる

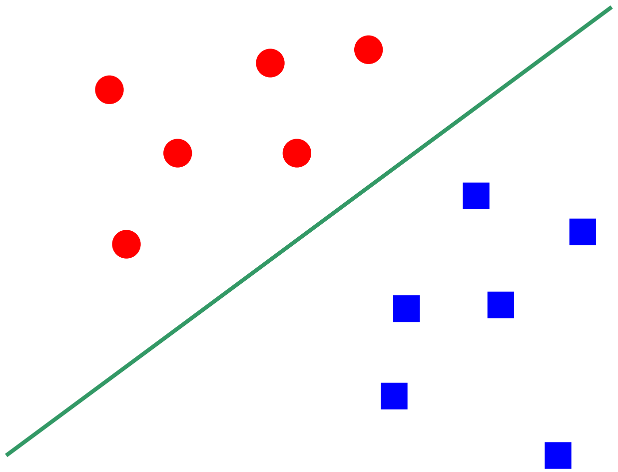
- RVM (Random vector machine)

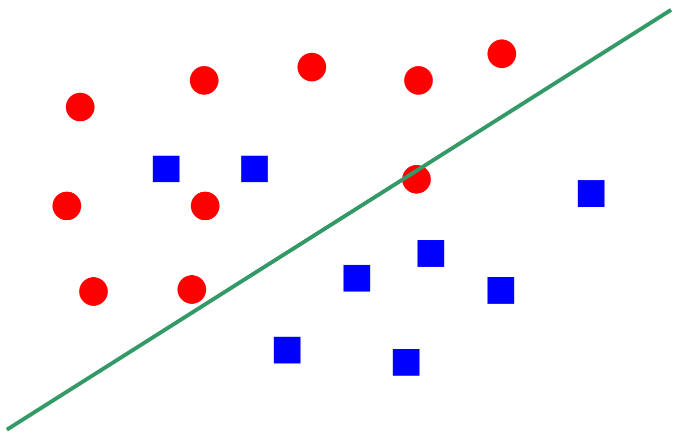
－ベイズ理論に基づき事後確率の推定が可能

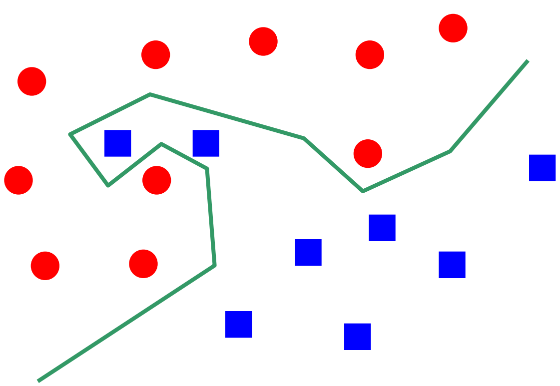
－SVMよりさらに疎なモデルが得られる

SVMの種類

- 線形SVM
  - ー線形分離可能
  - ー線形分離不可能
- 非線形SVM

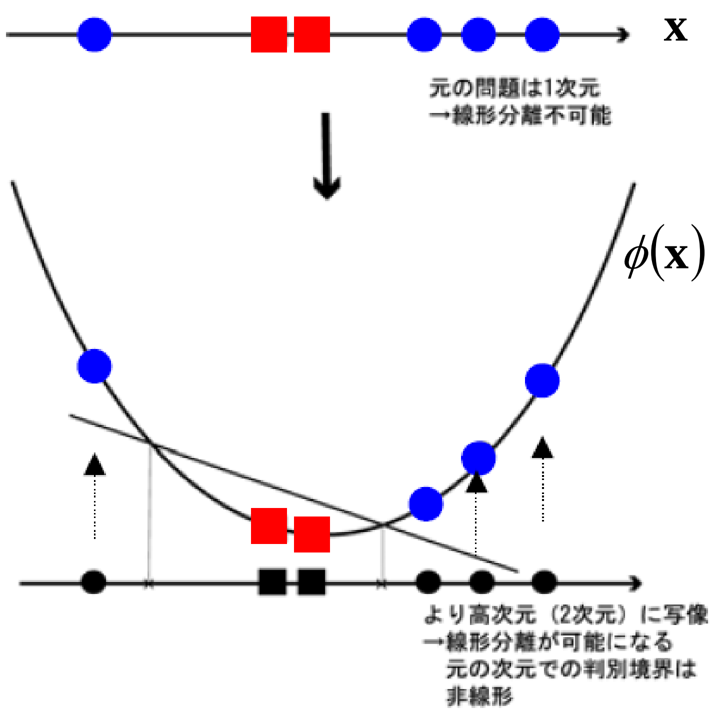






非線形SVM例

訓練データ点が特徴空間 $\phi(\mathbf{x})$ において線形分離可能
得られるSVMは入力空間 $\mathbf{x}$ において訓練データを
完全に分離する(入力空間においては分離境界は非線形にもなり得る)



クラス判別とSVM

２値分類(クラス分類)問題を解く

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b \quad (7.1)$$

$\phi(\mathbf{x})$:特徴空間変換関数

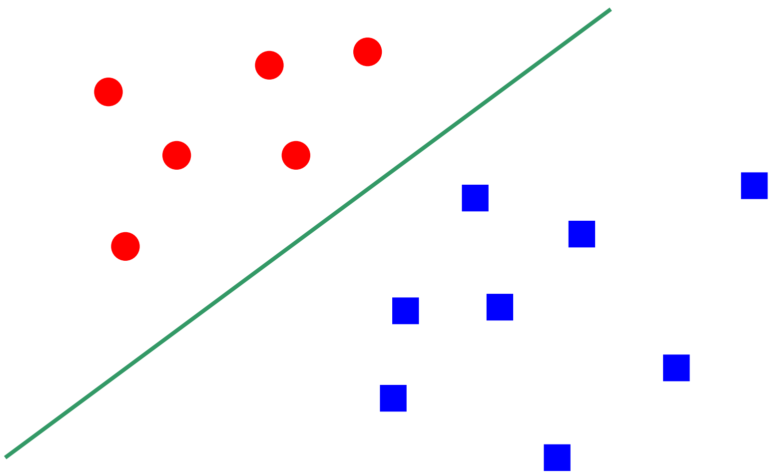
b :バイアスパラメータ

訓練データ　$\mathbf{x}_1, \ldots, \mathbf{x}_N$

目標値　$t_1, \ldots, t_N \; (t_n \in \{-1, 1\})$

未知のデータ点 $\mathbf{x}$ を

$y(\mathbf{x})$ の符号に応じて分類



最大マージン分類器

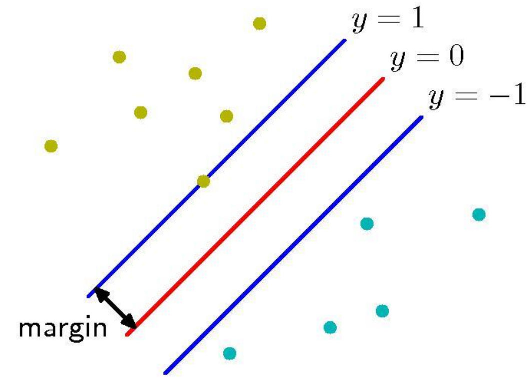
線形分離可能な場合

パラメータ $\mathbf{w}, b$ が存在

$$t_n = +1 \quad \forall x_n \ s.t\ y(\mathbf{x}_n) > 0$$

$$t_n = -1 \quad \forall x_n \ s.t\ y(\mathbf{x}_n) < 0$$

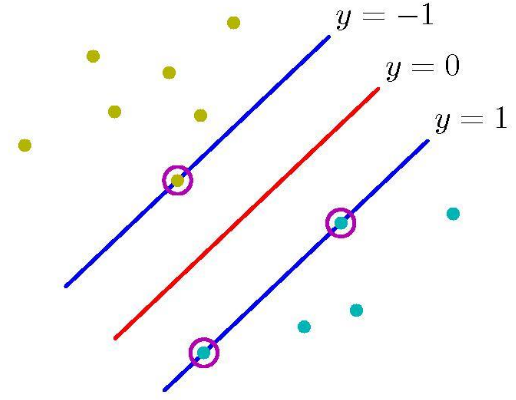
まとめて $t_n y(\mathbf{x}_n) > 0$



一般的には解は多数存在する

汎化誤差がもっとも小さくなるような解を求める

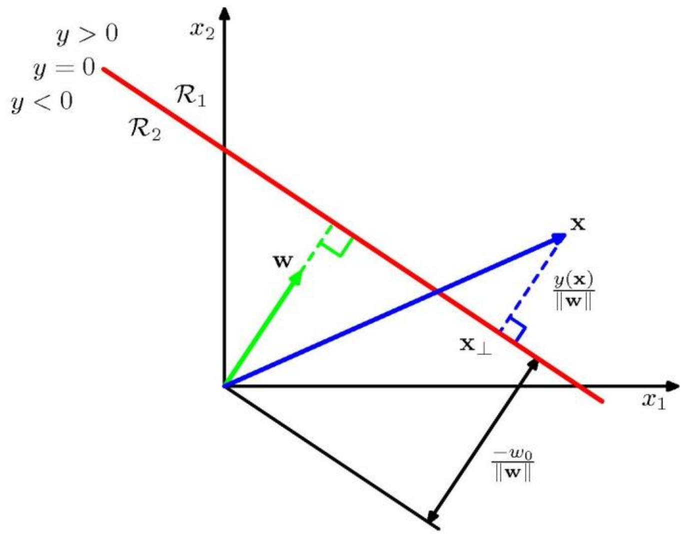
マージン(margin) を最大化する



マージン最大化定式化

超平面 $y(\mathbf{x})=0$ から点 $\mathbf{x}$ までの距離は $\dfrac{|y(\mathbf{x})|}{\|\mathbf{w}\|}$

線形分離可能の仮定から $t_n y(\mathbf{x})>0$



分解境界から点 $\mathbf{x}_n$までの距離は

$$\frac{t_n y(\mathbf{x_n})}{\|\mathbf{w}\|}=\frac{t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x_n})+b\right)}{\|\mathbf{w}\|} \quad (7.2)$$

マージンを最大化する最適化問題

$$\underset{\mathbf{w},b}{\arg\max}\left\{\frac{1}{\|\mathbf{w}\|}\min\left[t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)\right]\right\} \quad (7.3)$$

境界に最も近い点について $t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)=1 \quad (7.4)$

全ての点について $t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)\geq 1, \quad n=1,\ldots,N. \quad (7.5)$

マージン最大化問題

マージン最大化問題は以下の問題に帰着される

$$\underset{\mathbf{w},b}{\arg\min}\ \frac{1}{2}\|\mathbf{w}\|^2 \quad (7.6)$$

二次計画法の一例
→ 線形不等式系で与えられる制約条件の下で二次関数を最小化する問題

制約付き最適化問題を解くため，ラグランジュ乗数を導入する

$$L(\mathbf{w},b,\mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \{t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b)-1\} \quad (7.7)$$

$$\mathbf{a} = (a_1,\ldots,a_N)^T \quad a_n \geq 0$$

$\mathbf{w}, b$ に関する停留条件

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n) \quad (7.8) \qquad 0 = \sum_{n=1}^{N} a_n t_n \quad (7.9)$$

マージン最大化問題の双対表現

(7.6)の双対表現

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m). \quad (7.10)$$

以下のカーネル関数を用いた

$$k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m)$$

制約条件

$$a_n \geq 0, \quad n = 1, \ldots, N, \quad (7.11)$$

$$\sum_{n=1}^{N} a_n t_n = 0. \quad (7.12)$$

$\mathbf{a}$ に関して最大化

この問題は再び二次計画法になっている(最適化変数は $\mathbf{a}$)

双対表現

主問題(Primary problem)
M変数，N制約式

⇕

双対問題(Dual problem)
N変数，M制約式

カーネルトリック

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m). \quad (7.10)$$

カーネル関数

$$k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m)$$

カーネル関数が定義されていれば
$\phi(\mathbf{x})$ の具体的な形を知る必要はない

例） $k(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^t \mathbf{x}_m)^p$

$$\mathbf{x} = (x_1, x_2)^t$$

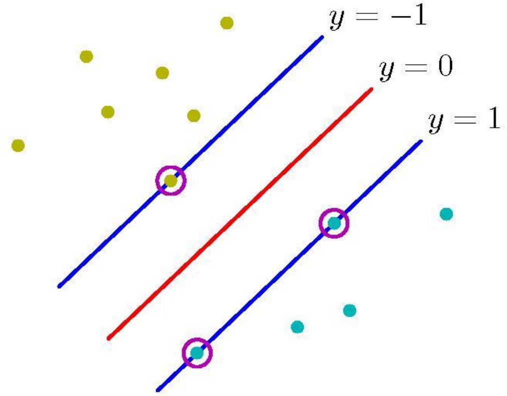
$$k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^t \phi(\mathbf{x}_m) = (1 + \mathbf{x}_n^t \mathbf{x}_m)^3$$

$$\phi(\mathbf{x}) = (1, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{6}x_1 x_2, \sqrt{3}x_1^2 x_2, \sqrt{3}x_1 x_2^2, x_1^3, x_2^3)^t$$

サポートベクトル

学習したモデルで新しいデータ点を分類するには

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b \quad (7.13)$$



このときKKT条件は

$$a_n \geq 0 \quad (7.14)$$

$$t_n y(\mathbf{x}_n) - 1 \geq 0 \quad (7.15)$$

$$a_n \{t_n y(\mathbf{x}_n) - 1\} = 0 \quad (7.16) \Leftrightarrow \quad a_n = 0 \quad or \quad t_n y(\mathbf{x}_n) = 1$$

$a_n \neq 0$ がデータ点の予測に影響 ⇨ サポートベクトル

一度モデルを学習してしまえばサポートベクトル以外の訓練データは不要

バイアスパラメータの算出

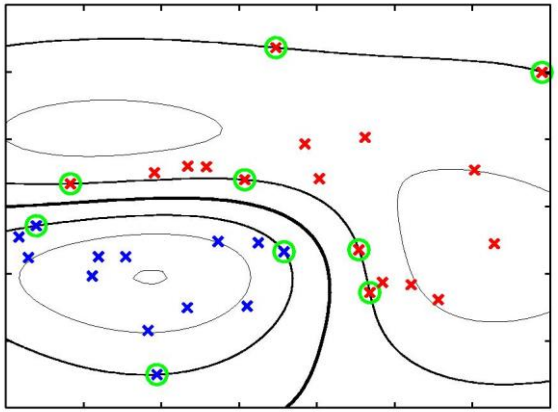
二次計画問題を解き $\mathbf{a}$ が求まるとバイアスパラメータ b を求めることができる

$$t_n\left(\sum_{m\in S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b\right) = 1 \quad (7.17)$$

S : サポートベクトルの集合

$$b = \frac{1}{N_S}\sum_{n\in S}\left(t_n - \sum_{m\in S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m)\right). \quad (7.18)$$

N_S : サポートベクトルの総数



線形分離不可能な場合への拡張

ここまでの仮定

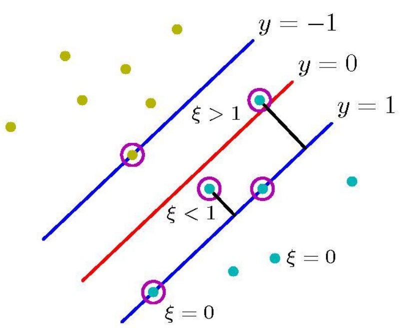
訓練データ点が特徴空間 $\phi(\mathbf{x})$ において線形分離可能
得られるSVMは入力空間 $\mathbf{x}$ において訓練データを
完全に分離する(入力空間においては分離境界は非線形にもなり得る)

⇩

訓練データを完全に分離する解が
必ずしも汎化能力に優れるとは限らない

⇩

一部の訓練データの誤分類を許すように修正



誤って分類したデータにはマージンの境界からの距離に比例した
ペナルティを与える

スラック変数　$\xi_n \geq 0 \,(n = 1, \ldots, N)$

データが正しく分類　$\xi_n = 0$　　　それ以外　$\xi_n = |t_n - y(\mathbf{x}_n)|$

ペナルティ項を加えた定式化

識別関数(7.5)を修正

$$t_n y(\mathbf{x}_n) \geq 1 - \xi_n, \quad n = 1, \ldots, N. \quad (7.20) \qquad \xi_n \geq 0 \; (n = 1, \ldots, N)$$

ハードマージンからソフトマージンへの緩和

目的はペナルティを与えつつもマージンを最大化する
よって次式を最小化する

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \|\mathbf{w}\|^2. \quad (7.21)$$

C :ペナルティとマージンの大きさのトレードオフを制御するパラメータ

$\sum \xi_n$:誤分類されたデータの上限

最小化問題のラグランジュ関数

ラグランジュ関数

$$L(\mathbf{w}, b, \xi, \mathbf{a}, \mu)$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}a_n\{t_n y(\mathbf{x}_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n \quad (7.22)$$

KKT条件

$$a_n \geq 0 \quad (7.23)$$

$$t_n y(\mathbf{x}_n) - 1 + \xi_n \geq 0 \quad (7.24)$$

$$a_n(t_n y(\mathbf{x}_n) - 1 + \xi_n) = 0 \quad (7.25)$$

$$\mu_n \geq 0 \quad (7.26)$$

$$\xi_n \geq 0 \quad (7.27)$$

$$\mu_n\xi_n = 0 \quad (7.28)$$

最小化問題の双対表現

$\mathbf{w}, b, \{\xi_n\}$ についての停留条件

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n) \quad (7.29)$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} a_n t_n = 0 \quad (7.30)$$

$$\frac{\partial L}{\partial \xi} = 0 \Rightarrow a_n = C - \mu_n \quad (7.31)$$

双対形のラグランジュ関数

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m). \quad (7.32)$$

制約条件

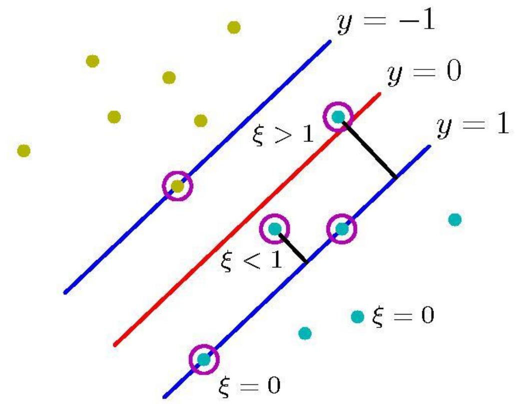
$$0 \le a_n \le C \quad (7.33)$$

$$\sum_{n=1}^{N} a_n t_n = 0 \quad (7.34)$$

サポートベクトル

サポートベクトルについては $a_n > 0$

$$t_n y(\mathbf{x}_n) = 1 - \xi_n \quad (7.35)$$



さらに $a_n \le C \Rightarrow \mu_n > 0 \quad \because (7.31)$

$$\mu_n > 0 \Rightarrow \xi_n = 0 \quad \because (7.28)$$

$0 < a_n < C$ のサポートベクトルでは $\xi_n = 0 \Rightarrow t_n y(\mathbf{x}_n) = 1$

先ほどと同様に $$t_n \left(\sum_{m \in S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b \right) = 1 \quad (7.36)$$

$$b = \frac{1}{N_M} \sum_{n \in M} \left(t_n - \sum_{m \in S} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right). \quad (7.37)$$

SVMを解くアルゴリズム

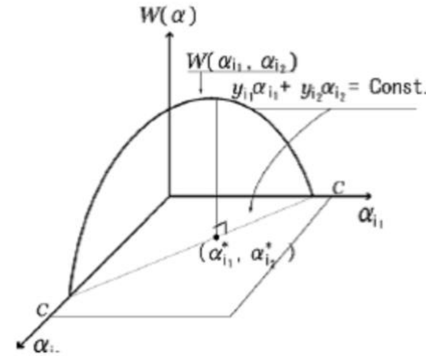
- 分類予測時と異なり，パラメータを学習する段階ではサポートベクトルでなく全ての訓練データの情報が必要
- 実用上，SVMの二次計画法を効率的に解くアルゴリズムが必要
- チャンギングを利用
- 保護共役勾配法(Burges, 1982)
- 分解法(Osuna *et al.*, 1996)
- 逐次最小問題最適化法(SMO；sequential minimal optimization)(Platt,1999)

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m). \quad (7.32)$$

SMOについて

- 計算効率(計算時間の飛躍的減少)のため，最もよく使われるSVMの最適化の手法
- たった２つのラグランジュ乗数を含む部分問題を逐次解いていくことで最終的な解を得る
- ２変数の選び方にはいくつかヒューリスティックが存在する
- 計算時間(データ数の１～２乗)
(一般的にデータの３乗)

SMOの定式



目的関数

$$\tilde{L}(a_{i1}, a_{i2}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

$$= a_{i_1} + a_{i_2} + f_1(a_{i_1}^2) + f_n(a_{i_2}^2) + f_3(a_{i_1}, a_{i_2})$$

$$+ f_4(a_{i_1}) + f_5(a_{i_2}) + \tilde{L}_{Const}.$$

制約条件

2変数の2次関数 ⇨ 1変数の2次関数

$$\sum_{n=1}^{N} a_n t_n = 0 \Leftrightarrow a_{i1} + a_{i2} = -\sum_{\substack{i=1 \\ i \neq i_1, i_2}}^{N} a_i t_i \quad (= Const)$$

すべての a_i がKTT条件を満たせば最適解となり終了

SMO続き

- 部分問題で取り上げる2変数の選び方

1. a_{i_1} を探索していき，KTT条件を満たしていない a_{i_1} を選択する．この時点で全ての a がKTT条件を満たす場合は，その時点で最適解となり，学習は終了となる．
2. ⅰ) a_{i_2} を選択する際に，まず $0 < a_{i_2} < C$ を満たし，かつ $|a_{i_2} - a_{i_1}|$ が最大になる a_{i_2} を選択する．
   ⅱ) $0 < a_{i_2} < C$ を満たす a_{i_2} がない場合には，$a_{i2} = C$ または $a_{i_2} = 0$ になる a_{i_2} を選択する．
   ⅲ) ⅰ)，ⅱ)がない場合，ランダムに a_{i_2} を選択する．
3. 選択された a_{i_1}, a_{i_2} で目的関数を解き，最適解が得られたら再び a_{i_1}, a_{i_2} を選択する．

SVMにまつわるその他のトピック

- ロジスティック回帰との関係
- 多クラスSVM
  - １対他(one-versus-the-rest)方式
  - １対１(one-versus-one)方式
- 回帰のためのSVM

SVMの問題点

- 多クラス識別が難しい
- ２次計画法を解くための計算量
- カーネルの選択

－カーネルの最適型

－カーネルの持つパラメータの最適値

－ペナルティとマージンの制御パラメータ

→実験で求める

何に使えるか

- BCALsの移動ー滞在判定

ラグランジュ乗数　その１

- 複数の変数に１つ以上の制約条件が課せられたときに，関数の停留点を求めるために用いられる．

例1） $\max\ f(x_1, x_2) \quad s.t.\ g(x_1, x_2) = 0$

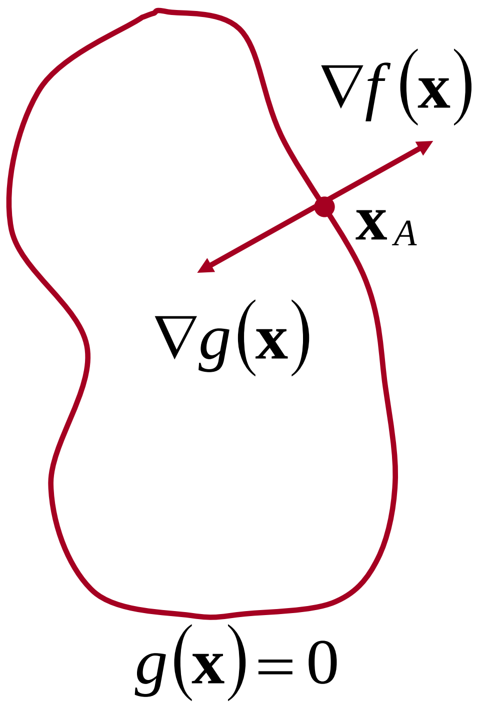
$$\nabla f + \lambda \nabla g = 0 \quad \lambda \neq 0 \quad (E.3)$$

ラグランジュ関数

$$L(\mathbf{x}, \lambda) \equiv f(\mathbf{x}) + \lambda g(\mathbf{x}) \quad (E.4)$$

$L(\mathbf{x}, \lambda)$ の $\mathbf{x}, \lambda$ に対する停留点を求める

$$\frac{\partial L}{\partial \mathbf{x}} = 0, \quad \frac{\partial L}{\partial \lambda} = 0$$



ラグランジュ乗数　その２

例2)　$\max f(x_1, x_2) \quad s.t.\ g(x_1, x_2) \geq 0$

ｉ）停留点が $g(\mathbf{x}) > 0$

制約が無効　⇨　停留条件　$\nabla f(\mathbf{x}) = 0$

$\lambda = 0$ の停留条件に等しい

ii）停留点が $g(\mathbf{x}) = 0$

解が制約面　$g(\mathbf{x}) = 0$ 上に存在　⇨　以前の停留条件に等しい

さらに　$\lambda > 0$ が存在して $\nabla f(\mathbf{x}) = -\lambda \nabla g(\mathbf{x})$

いずれにしろ　$\lambda g(\mathbf{x}) = 0$

ラグランジュ乗数　その３

例2） $\max\ f(x_1, x_2) \quad s.t.\ g(x_1, x_2) \geq 0$

以下の条件でラグランジュ関数(E.4)の停留点を求める

$$g(\mathbf{x}) \geq 0$$

$$\lambda \geq 0$$

$$\lambda g(\mathbf{x}) = 0$$

これを **Karush-Kuhn-Tucker**条件(KKT条件)という

何となく気になったこと