(83342) パターン認識特論

張　潮
zhang@u-fukui.ac.jp
0776-27-8477

本日の内容

第1回　パターン認識のための線形代数
第2回　パターン認識のための確率論
第3回　凸最適化の概念
第4回　線形回帰とLMSアルゴリズム
第5回　分類とロジスティック回帰
第6回　一般化線形モデル
第7回　ガウシアン判別分析と単純ベイズ分類器
第8回　サポートベクターマシン(1)
第9回　サポートベクターマシン(2)
第10回　正則化とモデル選択
第11回　k平均法
第12回　混合ガウスモデルとEMアルゴリズム
第13回　因子分析
第14回　主成分分析

第15回　未定

前回

- 学習データの幾何マージンを最大化することでよい超平面（決定境界）を見つけることができる
- マージンの最大化を凸最適化問題としてみなすことができる

$$
\begin{aligned}
\min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\
\text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m
\end{aligned}
$$

- 二次計画法で解けるが，全データを使うことになる（非効率的！）
- 一部のデータのみを使用し，最大なマージンを探す
  - 相対問題を利用する

前回

- カルーシュ・クーン・タッカー条件
  - Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*,\alpha^*,\cancel{\beta^*}) &= 0, \quad i=1,\ldots,n \\
\cancel{\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*,\alpha^*,\beta^*)} &\cancel{= 0, \quad i=1,\ldots,l} \\
\alpha_i^* g_i(w^*) &= 0, \quad i=1,\ldots,k \\
g_i(w^*) &\le 0, \quad i=1,\ldots,k \\
\alpha^* &\ge 0, \quad i=1,\ldots,k
\end{aligned}$$

SVMの場合

$$\begin{aligned}
\min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\
\text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \ge 1, \quad i=1,\ldots,m
\end{aligned}$$

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \le 0.$$

- KKT条件に満たす w^*, α^*, β^* は主問題と相対問題の解である

ラグランジュ相対性を利用した最適解の求め方

- wを求めたい

$$\mathcal{L}(w,\alpha,\beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \cancel{\sum_{i=1}^{l} \beta_i h_i(w)}.$$

$\alpha_i = 0$の項は、最適解wに影響しない

$$\begin{array}{rcl}
\dfrac{\partial}{\partial w_i}\mathcal{L}(w^*,\alpha^*,\cancel{\beta^*}) & = & 0,\quad i = 1,\ldots,n \\
\cancel{\dfrac{\partial}{\partial \beta_i}\mathcal{L}(w^*,\alpha^*,\beta^*)} & \cancel{=} & \cancel{0,\quad i = 1,\ldots,l} \\
\boxed{\alpha_i^*} \cancel{g_i(w^*)} & \cancel{=} & \cancel{0,\quad i = 1,\ldots,k} \\
g_i(w^*) & \le & 0,\quad i = 1,\ldots,k \\
\alpha^* & \ge & 0,\quad i = 1,\ldots,k
\end{array}$$

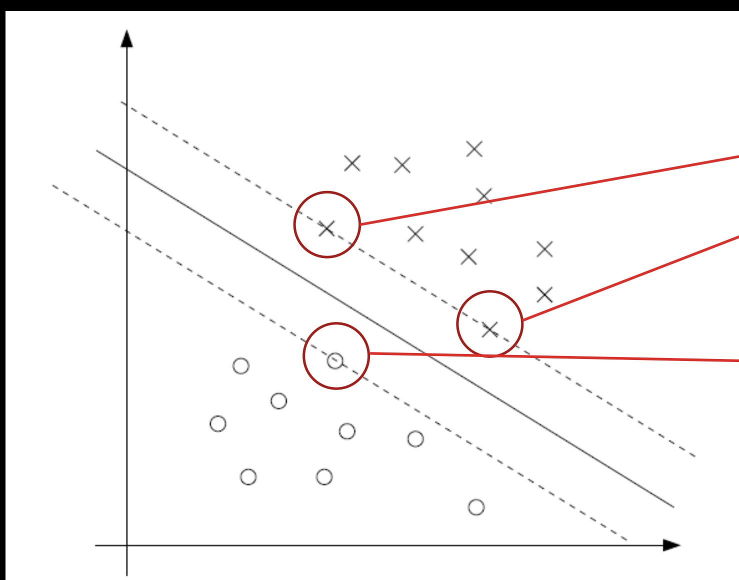
>0なら$g_i(w)=0$

$$g_i(w) = -\underbrace{y^{(i)}(w^T x^{(i)} + b)}_{} + 1 \le 0.$$

|

ラグランジュ相対性を利用した最適解の求め方

$$g_i(w) = -\underbrace{y^{(i)}(w^T x^{(i)} + b)}_{\mathrm{I}} + 1 \leq 0.$$



$y^{(i)} = 1,\ w^T x^{(i)} + b = 1$

$y^{(i)} = -1,\ w^T x^{(i)} + b = -1$

サポートベクトル！！

ラグランジュ相対性を利用した最適解の求め方

- g_iを代入すると

$$\mathcal{L}(w,b,\alpha)=\frac{1}{2}||w||^2-\sum_{i=1}^{m}\alpha_i\left[y^{(i)}(w^Tx^{(i)}+b)-1\right].$$

- wとbにおける微分=0で*L*を最小化

$$\nabla_w\mathcal{L}(w,b,\alpha)=w-\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}=0$$

$$w=\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}.$$

$$\frac{\partial}{\partial b}\mathcal{L}(w,b,\alpha)=\sum_{i=1}^{m}\alpha_i y^{(i)}=0.$$

ラグランジュ相対性を利用した最適解の求め方

- 微分=0の解を代入することで最小化

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}.$$

$$\frac{\partial}{\partial b}\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

代入　　制限

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right].$$

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} - b\sum_{i=1}^{m} \alpha_i y^{(i)}.$$

ラグランジュ相対性を利用した最適解の求め方

- max min で相対問題を作る

$$
\begin{aligned}
\max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle. \\
\text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \\
& \sum_{i=1}^{m} \alpha_i y^{(i)} = 0,
\end{aligned}
$$

αが解ける（解き方については後で紹介する）→wが分かる→bが分かる

$$
b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}}{2}.
$$

SVMでの予測

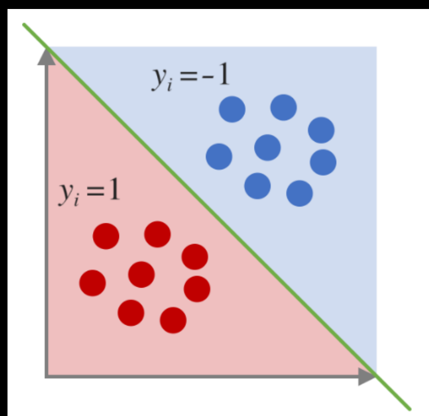
- 最適化されたパラメータで新しいテストデータを予測
  - 入力:x
  - 下記の式が>0ならy=1と予測する
  - 下記の式が<0ならy=-1と予測する

=0の項を無視できる

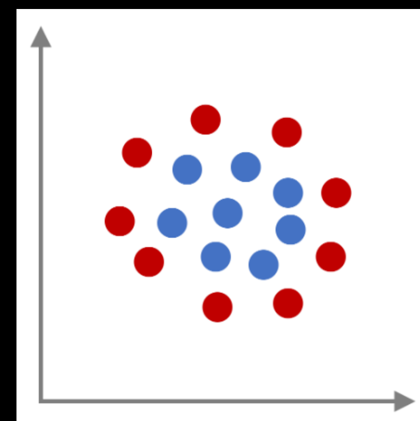
$$
\begin{aligned}
w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x + b \\
&= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.
\end{aligned}
$$

線形分離不可能なデータに対応できない☹

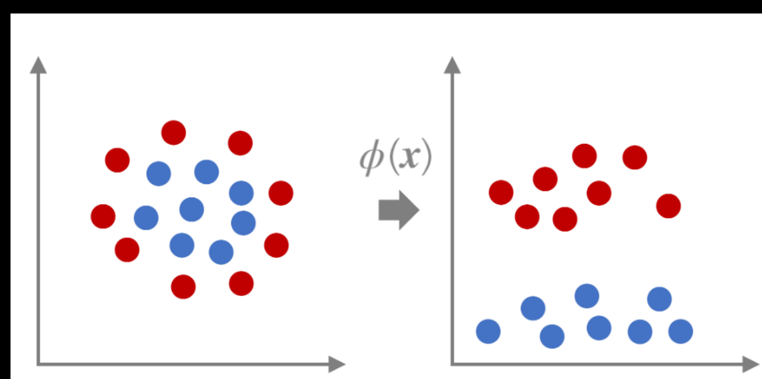
②カーネル関数（kernel function）



☺



☹



☺

②カーネル関数（kernel function）

- 属性とは？特徴量とは？

属性(attributes)

特徴量マッピング
(feature mapping)

特徴量(features)

$$\phi(x) = \begin{bmatrix} x \\ x^2 \\ x^3 \end{bmatrix}.$$

- 属性ではなく、特徴量で学習することもできる

②カーネル関数(kernel function)

$$w^T x + b = \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x + b$$
$$= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.$$

$\langle x, z \rangle \rightarrow \langle \phi(x), \phi(z) \rangle$

属性の代わりに、特徴量を使う

$$K(x, z) = \phi(x)^T \phi(z).$$

カーネル:特徴量ベクトルの内積

②カーネル関数（kernel function）

- $$K(x,z) = (x^T z)^2.$$
- $$K(x,z) = \phi(x)^T \phi(z).$$ の形で書くことが可能

■この場合の特徴量マッピング関数Φは？

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{n} x_i z_i\right)\left(\sum_{j=1}^{n} x_i z_i\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{n} (x_i x_j)(z_i z_j) \end{aligned}$$

n=3の場合

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}.$$

②カーネル関数（kernel function）

- Φ(x)とΦ(z) が似ていればK(x,z)が小さくなる
  - 直交した場合は0になる

$$K(x, z) = \phi(x)^T \phi(z).$$

- カーネルをガウシャンカーネルにしたらΦはどうなる？

$$K(x, z) = \exp\left(-\frac{||x - z||^2}{2\sigma^2}\right).$$

Φ(x)が無限次元になる

②カーネル関数（kernel function）

- 任意のカーネルが有効というわけではない
  - 対応しているマッピング関数が存在しない場合も
  - マッピング関数の存在を判定する必要がある

  - カーネル行列で判定を行う

$$K_{ij} = K(x^{(i)}, x^{(j)})$$

$$K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)}) = \phi(x^{(j)})^T \phi(x^{(i)}) = K(x^{(j)}, x^{(i)}) = K_{ji}$$

対称行列である

②カーネル関数（kernel function）

● カーネル行列の二次形式を見てみよう

$$
\begin{aligned}
z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j \\
&= \sum_i \sum_j z_i \sum_k \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \sum_i \sum_j z_i \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \left(\sum_i z_i \phi_k(x^{(i)}) \right)^2 \\
&\geq 0.
\end{aligned}
$$

カーネル行列が対称半正定値行列なら、
カーネルの特徴マッピング関数Φが存在し、
カーネルが有効である。
（実は、必要条件だけではなく、必要十分条件である）

半正定値行列である（1回目授業の24ページ）

②カーネル関数（kernel function）

● Mercerの定理

Theorem (Mercer). Let $K : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ be given. Then for K to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\{x^{(1)}, \ldots, x^{(m)}\}$, $(m < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.

- Mercer, J. (1909), "Functions of positive and negative type and their connection with the theory of integral equations", Philosophical Transactions of the Royal Society A, 209 (441–458): 415&ndash, 446, doi:10.1098/rsta.1909.0016

②カーネル関数（kernel function）

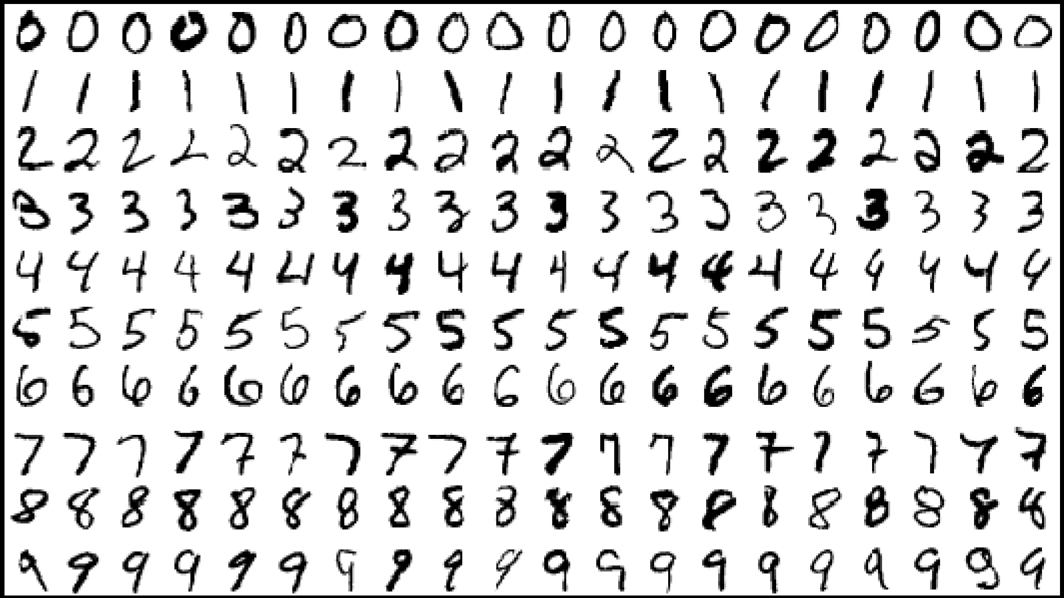
- カーネルという考え方は、SVM以外でもある
- 属性ベクトルの内積をカーネルベクトルの内積で書き換えるだけで、高次元や線形分離不可のデータに対応できるようになる

⬇

カーネルトリック（kernel trick）

カーネルを取り入れたSVMの例

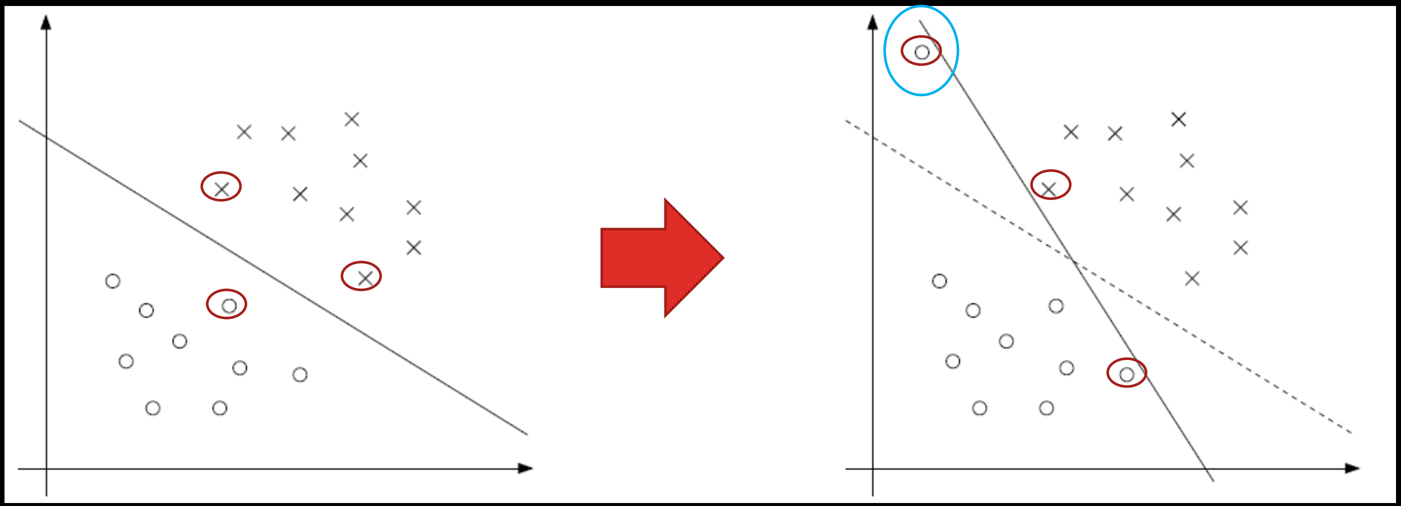
- 手書き数字の認識
  - 16ピクセル×16ピクセル
  - 多項式カーネルorガウシャンカーネルが有効



正則化と外れ値への対応

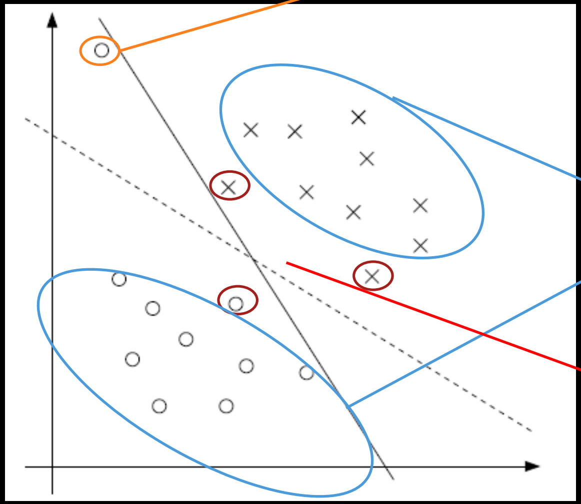
- 外れ値
  - 関数マージン=1の点のみ使われているため

$$y^{(i)}(w^T x^{(i)} + b) = 1$$



正則化と外れ値への対応

$y^{(i)}(w^T x^{(i)} + b) \leq 1$ 使う



$y^{(i)}(w^T x^{(i)} + b) \geq 1$ 使わない

$y^{(i)}(w^T x^{(i)} + b) = 1$ 使う

外れ値をペナルティとして考える

正則化と外れ値への対応

- 条件を緩める
- ペナルティ項を追加

$$
\begin{aligned}
\min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\
\text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m
\end{aligned}
$$

$$
\begin{aligned}
\min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\
\text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\
& \xi_i \geq 0, \quad i = 1, \ldots, m.
\end{aligned}
$$

正則化と外れ値への対応

- 正則化項を追加した後のKKT条件の一部

追加前の変わらない

$$
\begin{aligned}
&\boxed{\alpha_i = 0 \quad \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) \geq 1} \\
&\alpha_i = C \quad \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) \leq 1 \\
0 < &\alpha_i < C \quad \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) = 1.
\end{aligned}
$$

④逐次最小問題最適化法（Sequential Minimal Optimization, SMO）

- 残りの最後の問題

$$
\begin{aligned}
\max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle. \\
\text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \\
& \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.
\end{aligned}
$$

④逐次最小問題最適化法（Sequential Minimal Optimization, SMO）

- 座標降下法
  - Coordinate Descent Algorithm

$$\max_{\alpha} W(\alpha_1, \alpha_2, \ldots, \alpha_m).$$

Loop until convergence: {

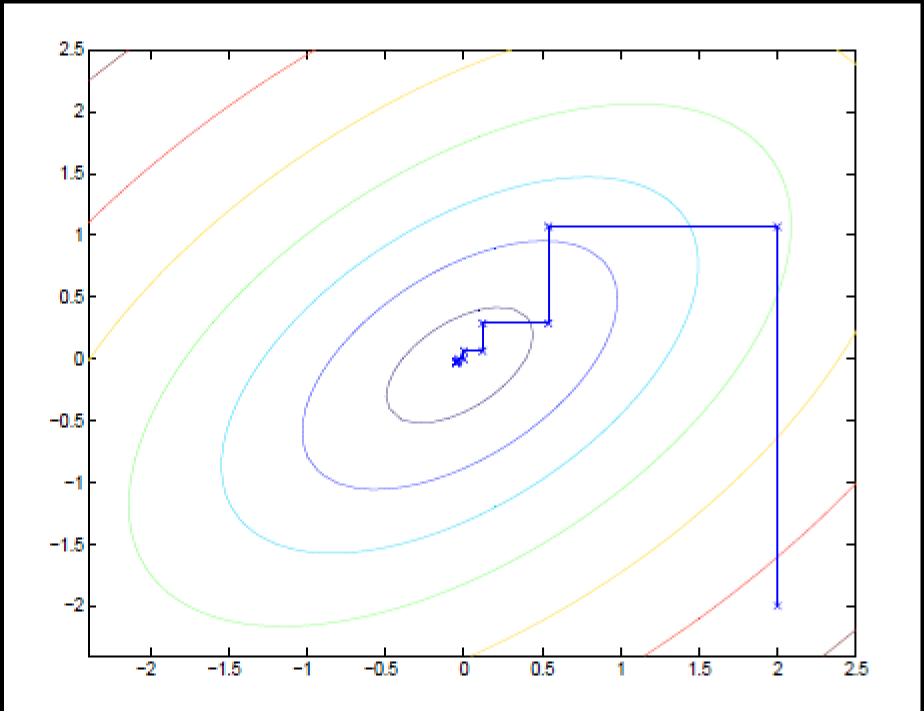
For $i = 1, \ldots, m$, {

$\alpha_i := \arg\max_{\hat{\alpha}_i} W(\alpha_1, \ldots, \alpha_{i-1}, \hat{\alpha}_i, \alpha_{i+1}, \ldots, \alpha_m).$

}

}

④逐次最小問題最適化法（Sequential Minimal Optimization, SMO）



④逐次最小問題最適化法（Sequential Minimal Optimization, SMO）

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle.$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

$$\alpha_1 y^{(1)} = -\sum_{i=2}^{m} \alpha_i y^{(i)}.$$

$$\alpha_1 = -y^{(1)} \sum_{i=2}^{m} \alpha_i y^{(i)}.$$

a1はa2～amで決まるため、a2～amを固定すると、a1を変更することができない

④逐次最小問題最適化法（Sequential Minimal Optimization, SMO）

- One-by-one できないなら、two-by-twoでやろう

Repeat till convergence {

1. Select some pair α_i and α_j to update next (using a heuristic that tries to pick the two that will allow us to make the biggest progress towards the global maximum).
2. Reoptimize $W(\alpha)$ with respect to α_i and α_j, while holding all the other α_k's ($k \neq i, j$) fixed.

}

④逐次最小問題最適化法（Sequential Minimal Optimization, SMO）

- α3～αmを固定し、α1およびα2を最適化すると考えよう

$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = -\sum_{i=3}^{m} \alpha_i y^{(i)}.$$

固定のため、定数ζで表す

$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = \zeta.$$

$$\alpha_1 = (\zeta - \alpha_2 y^{(2)}) y^{(1)}.$$

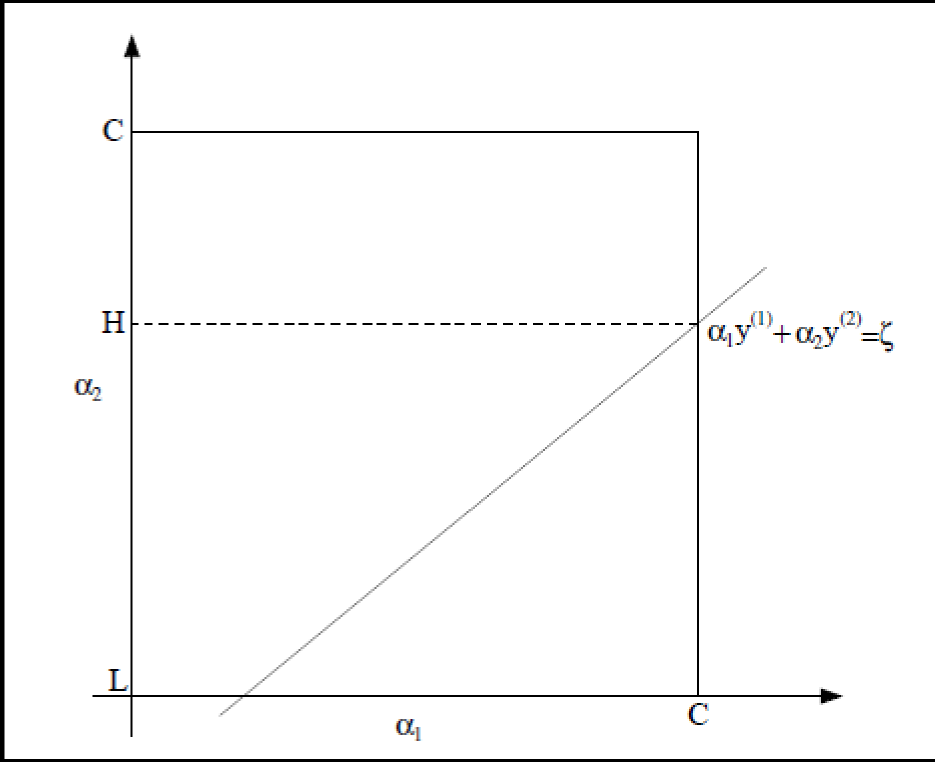
代入

$$W(\alpha_1, \alpha_2, \ldots, \alpha_m) = W((\zeta - \alpha_2 y^{(2)}) y^{(1)}, \alpha_2, \ldots, \alpha_m).$$

α2に関する多項式になった
（微分=0で最適化できる）

④逐次最小問題最適化法（Sequential Minimal Optimization, SMO）

- α2に範囲制限あり



④逐次最小問題最適化法（Sequential Minimal Optimization, SMO）

- 範囲からはみ出したときに”クリップ”処理を行う

$$
\alpha_2^{new} = \begin{cases} H & \text{if } \alpha_2^{new,unclipped} > H \\ \alpha_2^{new,unclipped} & \text{if } L \le \alpha_2^{new,unclipped} \le H \\ L & \text{if } \alpha_2^{new,unclipped} < L \end{cases}
$$

次回は

第1回　パターン認識のための線形代数
第2回　パターン認識のための確率論
第3回　凸最適化の概念
第4回　線形回帰とLMSアルゴリズム
第5回　分類とロジスティック回帰
第6回　一般化線形モデル
第7回　ガウシアン判別分析と単純ベイズ分類器
第8回　サポートベクターマシン(1)
第9回　サポートベクターマシン(2)
第10回　正則化とモデル選択
第11回　k平均法
第12回　混合ガウスモデルとEMアルゴリズム
第13回　因子分析
第14回　主成分分析
第15回　未定