Support Vector Machine

参考文献

- C. Cortes and V. **Vapnik**, *Support-Vector Networks*, Machine Learning, 20(3):273-297, September 1995
- Vladimir N. **Vapnik**. *The Nature of Statistical Learning Theory*. Springer, New York, 1995

参考サイト http://www.kernel-machines.org/

参考 java applet http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Bio-informatics への応用

- **マイクロアレーデータから遺伝子や組織を分類**

  Knowledge-based analysis of microarray gene expression data by using support vector machines, Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terence S. Furey, Manuel Ares, Jr., David Haussler, Proc. Natl. Acad. Sci. USA, vol. 97, pages 262-267

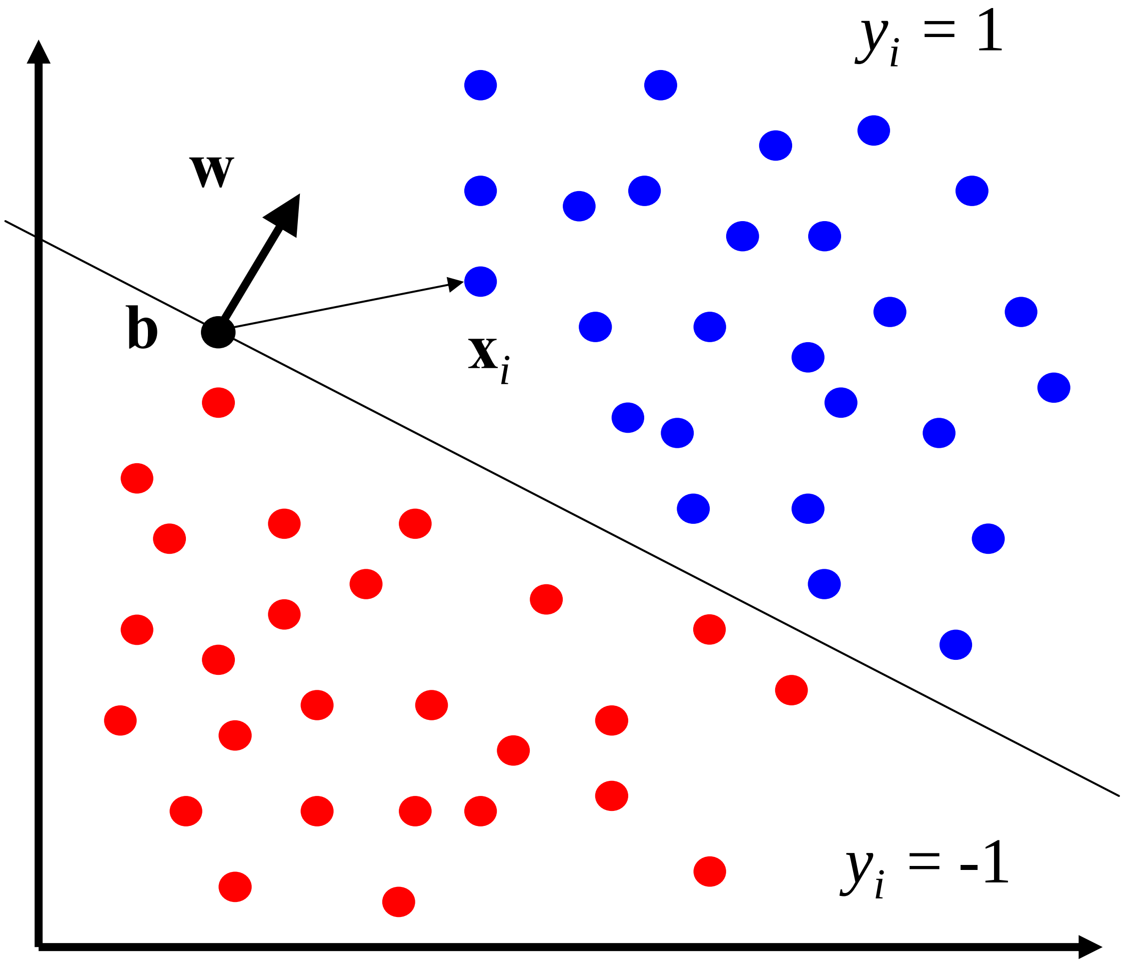
  Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, Terrence S. Furey, Nigel Duffy, Nello Cristianini, David Bednarski, Michel Schummer, and David Haussler, Bioinformatics. 2000, 16(10):906-914.

- **タンパク質の結合関係の予測**

  Multi-class protein fold recognition using support vector machines and neural networks, Chris Ding and Inna Dubchak, Bioinformatics, 17:349-358, 2001

  Predicting protein-protein interactions from primary structure w, Joel R. Bock and David A. Gough, Bioinformatics 2001 17: 455-460

Fisher の線形判別



訓練データ

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l) \quad \mathbf{x}_i \in R^n, y_i \in \{-1, +1\}$$

が線形分離可能とは、
ベクトル $\mathbf{w}$ と $\mathbf{b}$ が存在し

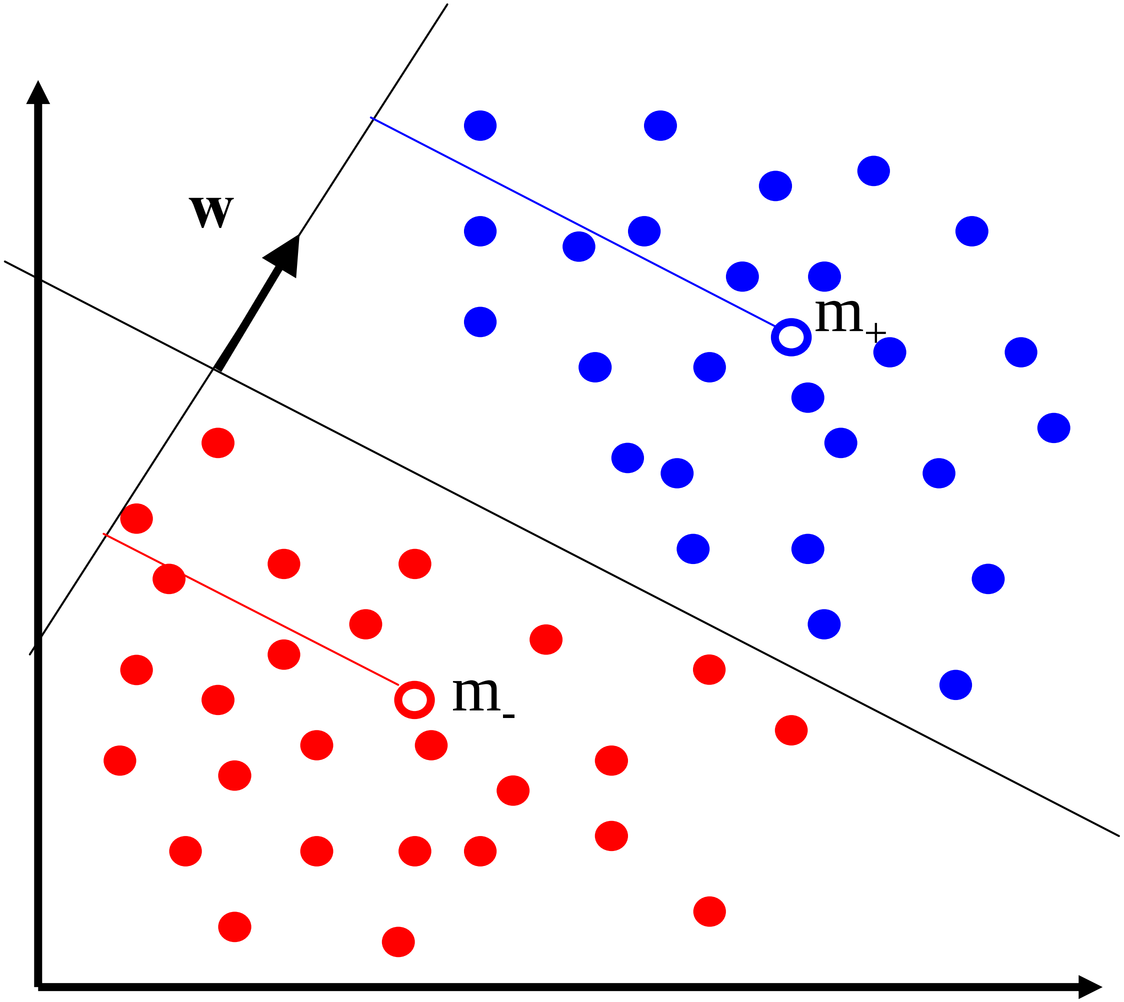
$$y_i(\mathbf{w} \cdot (\mathbf{x}_i - \mathbf{b})) > 0 \qquad (i = 1, \dots, l)$$

$\mathbf{b}$ の自由度は不必要に多い
$-\mathbf{w} \cdot \mathbf{b}$ を定数 b で表現すれば、

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 0 \qquad (i = 1, \dots, l)$$

を満たす $\mathbf{w}$ と b が存在することと同値

Fisher の線形判別



線形判別関数をつかっ　たクラス分類 $d(\mathbf{x})$

$$d(\mathbf{x}) = \begin{cases} 1 & \mathbf{w}\cdot\mathbf{x} + b \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

正および負に分類され　たデータの平均値

$$\mathbf{m}_{+} = \frac{\sum_{d(\mathbf{x})=1} \mathbf{x}}{\left|\{\mathbf{x} \mid d(\mathbf{x}) = 1\}\right|} \quad \mathbf{m}_{-} = \frac{\sum_{d(\mathbf{x})=-1} \mathbf{x}}{\left|\{\mathbf{x} \mid d(\mathbf{x}) = -1\}\right|}$$

$\mathbf{m}_{+}, \mathbf{m}_{-}$ を $\mathbf{w}$ 上に射影した点の距離　の二乗

$$((\mathbf{m}_{+} - \mathbf{m}_{-})\cdot\mathbf{w})^2$$

を大きく、正と負のク　ラス内での点の分散

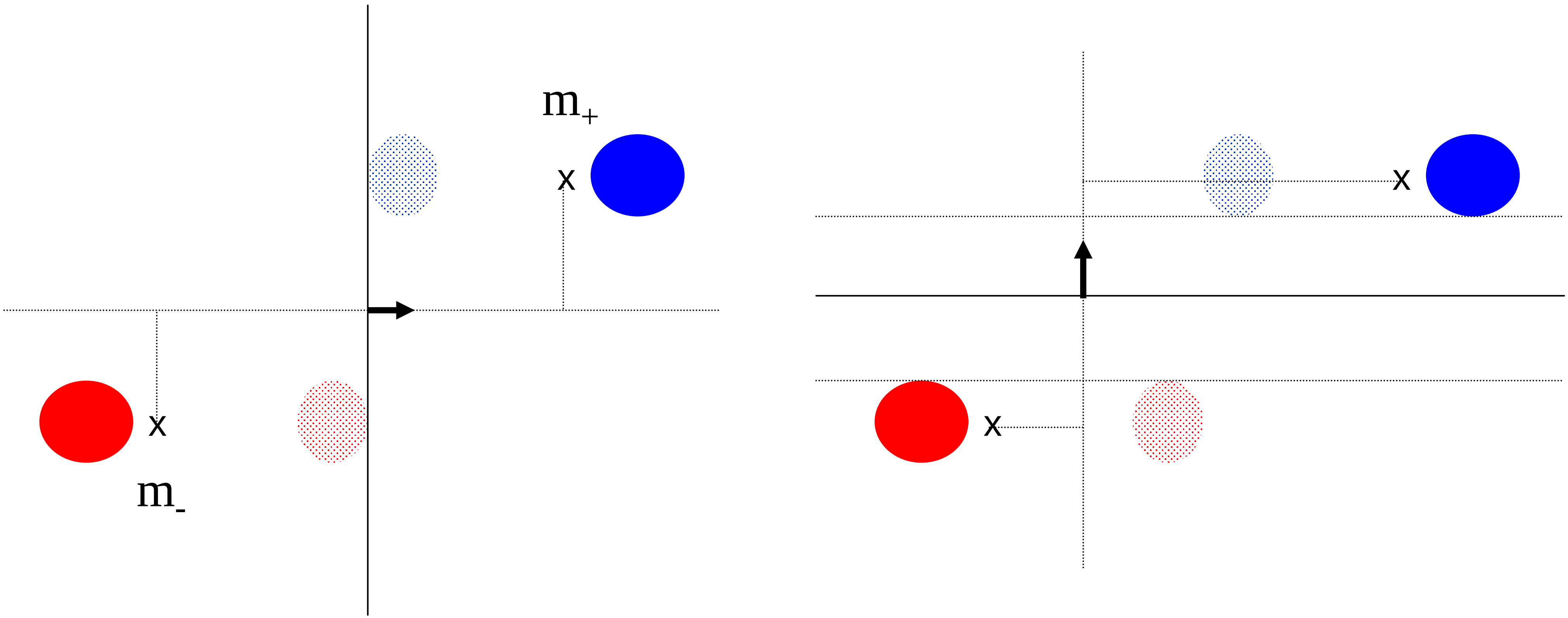
$$\sum_{d(\mathbf{x})=1} ((\mathbf{x} - \mathbf{m}_{+})\cdot\mathbf{w})^2 + \sum_{d(\mathbf{x})=-1} ((\mathbf{x} - \mathbf{m}_{-})\cdot\mathbf{w})^2$$

を小さくしたい．　そこで

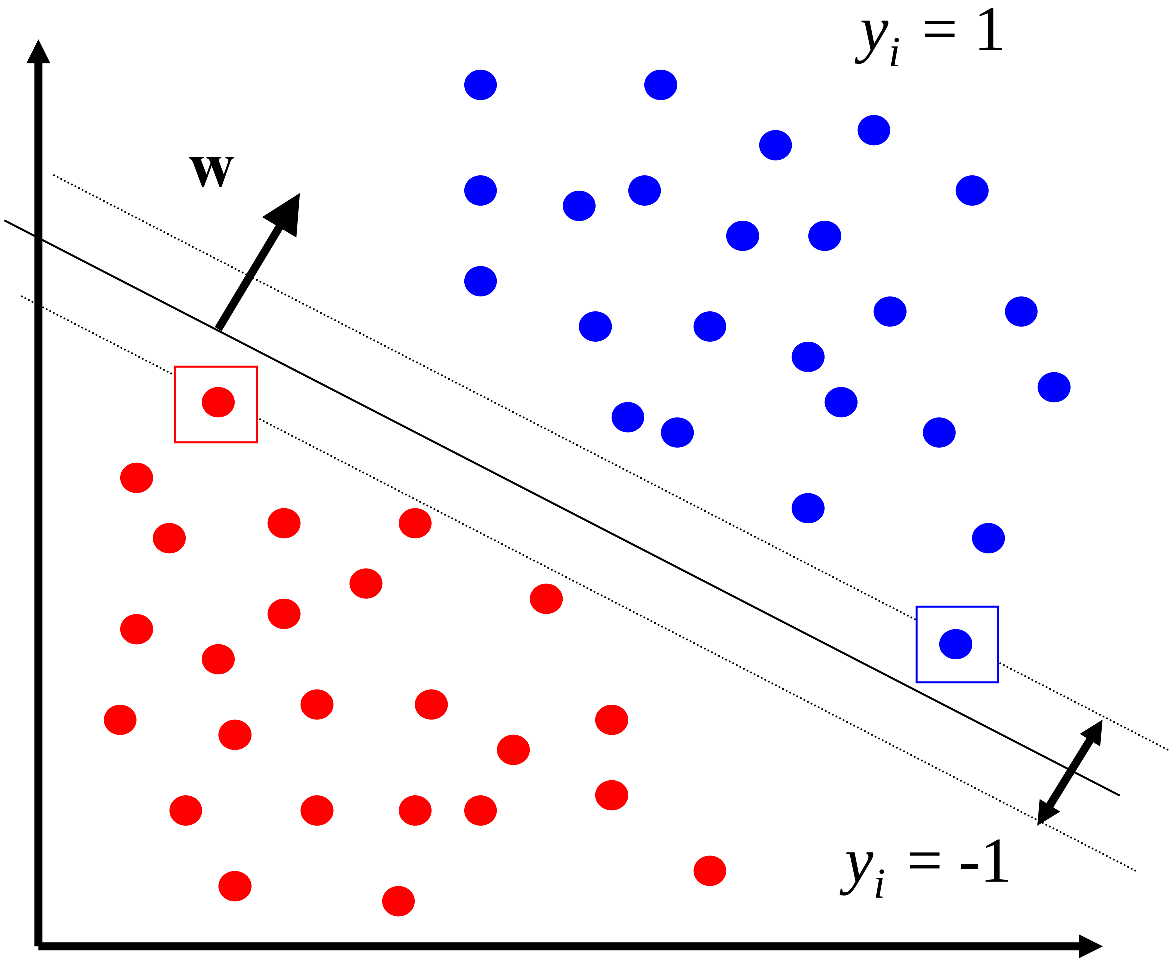
$$\frac{((\mathbf{m}_{+} - \mathbf{m}_{-})\cdot\mathbf{w})^2}{\sum_{d(\mathbf{x})=1} ((\mathbf{x} - \mathbf{m}_{+})\cdot\mathbf{w})^2 + \sum_{d(\mathbf{x})=-1} ((\mathbf{x} - \mathbf{m}_{-})\cdot\mathbf{w})^2}$$

を最大にする　$\mathbf{w}$ を求める

線形判別とマージン



Support Vector Machine
マージン (Margin)

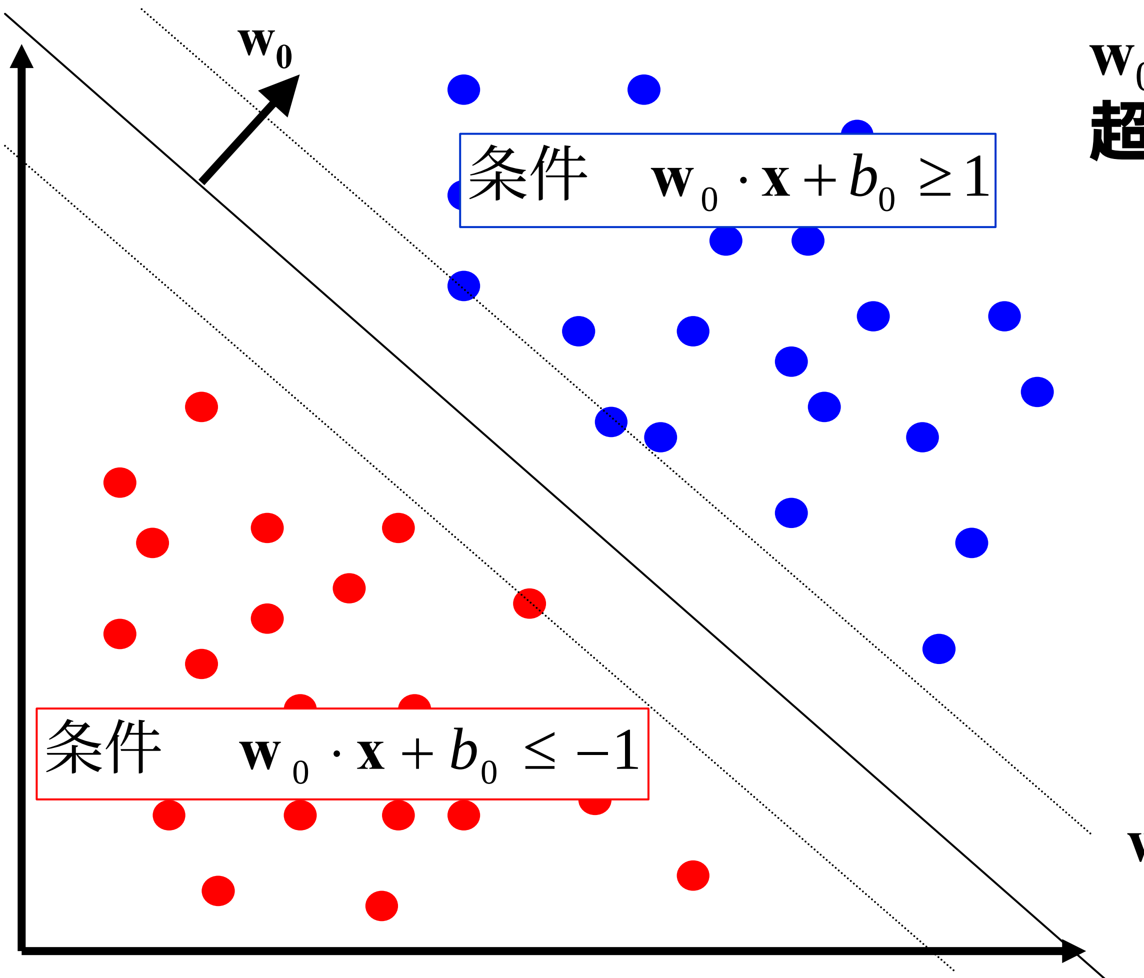


SVM **はマージンを最大化するアプローチ**

マージン境界上の点
(support vector) **に依存**

$$\text{マージン} \quad \rho(\mathbf{w}, b) = \min_{\{\mathbf{x}_i \mid y_i = 1\}} \frac{\mathbf{x}_i \cdot \mathbf{w}}{|\mathbf{w}|} - \max_{\{\mathbf{x}_i \mid y_i = -1\}} \frac{\mathbf{x}_i \cdot \mathbf{w}}{|\mathbf{w}|}$$

最適超平面 (Optimal Hyperplanes)



$\mathbf{w}_0$ がマージンを最大化する超平面を与える場合を考えると

$$
\begin{aligned}
&\rho(\mathbf{w}_0, b_0) \\
&= \min_{\{\mathbf{x}_i \mid y_i = 1\}} \frac{\mathbf{x}_i \cdot \mathbf{w}_0}{|\mathbf{w}_0|} - \max_{\{\mathbf{x}_i \mid y_i = -1\}} \frac{\mathbf{x}_i \cdot \mathbf{w}_0}{|\mathbf{w}_0|} \\
&= \frac{1 - b_0}{|\mathbf{w}_0|} - \frac{-1 - b_0}{|\mathbf{w}_0|} = \frac{2}{|\mathbf{w}_0|}
\end{aligned}
$$

境界線上に載る点が存在するように $\mathbf{w}_0$ の大きさを調節

最適超平面：2次計画問題へ

[2次計画問題]

制約　$y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1$　　$(i=1,\ldots,l)$　のもと　$2/\sqrt{\mathbf{w}\cdot\mathbf{w}}$ を最大化、

つまり $\mathbf{w}\cdot\mathbf{w}$ を最小化する $\mathbf{w}_0$ が最適超平面をあたえる

ラグランジュ法により解く

$\mathbf{\Lambda}^T=(\alpha_1,\ldots,\alpha_l)$ をラグランジュ乗数 $\alpha_i\ (\geq 0)$

$$L(\mathbf{w},b,\mathbf{\Lambda})=\frac{1}{2}\mathbf{w}\cdot\mathbf{w}-\sum_{i=1}^{l}\alpha_i\left[y_i(\mathbf{x}_i\cdot\mathbf{w}+b)-1\right]\quad\cdots\quad(1)$$

$\mathbf{w}$ と b について最小、$\mathbf{\Lambda}$ について最大となる saddle point(鞍点) が $\mathbf{w}_0$ を与える

最適超平面：双対問題へ

$\mathbf{w}, b$ について最小であることの必要条件

$$\left.\frac{\partial L(\mathbf{w}, b, \mathbf{\Lambda})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}_0} = \mathbf{w}_0 - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = 0 \qquad \mathbf{w}_0 = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \quad \cdots \quad (2)$$

$$\left.\frac{\partial L(\mathbf{w}, b, \mathbf{\Lambda})}{\partial b}\right|_{b=b_0} = -\sum_{i=1}^{l} \alpha_i y_i = 0$$

(1) へ代入

$$L(\mathbf{w}_0, b_0, \mathbf{\Lambda})$$

$$= \frac{1}{2}\mathbf{w}_0 \cdot \mathbf{w}_0 - \sum_{i=1}^{l} \alpha_i \left[y_i (\mathbf{x}_i \cdot \mathbf{w}_0 + b_0) - 1 \right]$$

$$= \frac{1}{2}\mathbf{w}_0 \cdot \mathbf{w}_0 - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{w}_0 - \sum_{i=1}^{l} \alpha_i y_i b_0 + \sum_{i=1}^{l} \alpha_i$$

$$= \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\mathbf{w}_0 \cdot \mathbf{w}_0 = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

[双対問題]

制約 $\sum_{i=1}^{l} \alpha_i y_i = 0,\ \alpha_i \geq 0$ のもと

$$L(\mathbf{w}_0, b_0, \mathbf{\Lambda}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

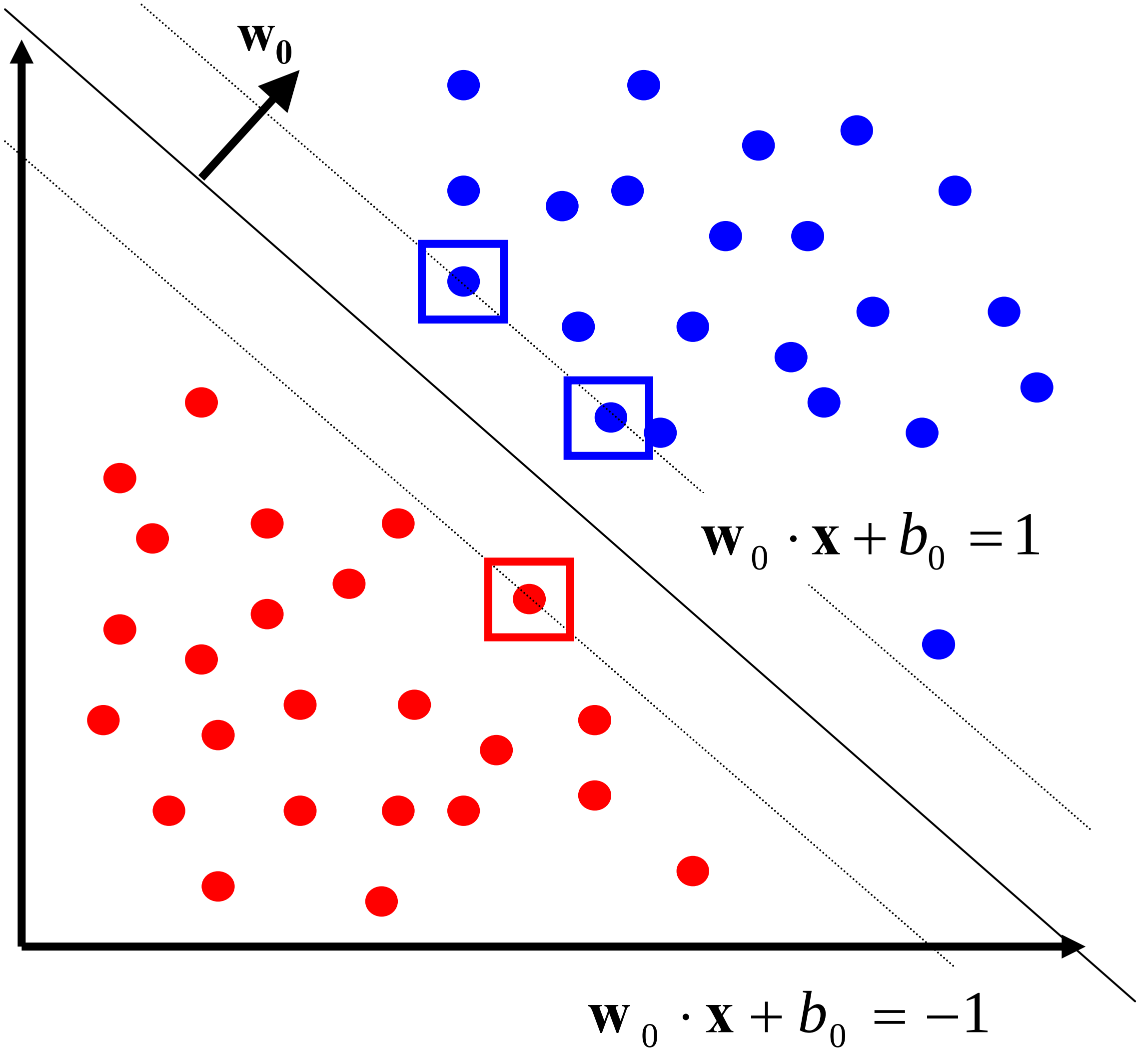
を最大化する $\mathbf{\Lambda}^T = (\alpha_1, \ldots, \alpha_l)$ を

2 次計画法で計算.

(2) $\mathbf{w}_0 = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$ より α_i から $\mathbf{w}_0$ を決定

すべての $\mathbf{x}_i$ が必要なのか?

最適超平面： $\mathbf{w}_0$ の計算



(2) より $\mathbf{w}_0$ は $\mathbf{x}_i$ の線形結合 $\sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$

すべての $\mathbf{x}_i$ が必要ではなくて

$\alpha_i \neq 0$ のときの $\mathbf{x}_i$ だけが線形結合に貢献

Kuhn - Tucker の定理より、

$\mathbf{w} \cdot \mathbf{w}$ を最小化する鞍点において

$$\alpha_i \left[y_i \left(\mathbf{x}_i \cdot \mathbf{w}_0 + b_0 \right) - 1 \right] = 0 \qquad (i = 1, \ldots, l)$$

$\alpha_i \neq 0$ のとき

$$y_i \left(\mathbf{x}_i \cdot \mathbf{w}_0 + b_0 \right) = 1$$

この条件を満たす $\mathbf{x}_i$ を support vector

<u>最適超平面は support vector で表現可能</u>

最適超平面: マージンの計算

$\boldsymbol{\Lambda}_0 = \left(\alpha_1^0, \ldots, \alpha_l^0\right)$ が $L(\mathbf{w}_0, b_0, \boldsymbol{\Lambda}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \mathbf{w}_0 \cdot \mathbf{w}_0$ を最大化すると仮定.

$$\mathbf{w}_0 \cdot \mathbf{w}_0$$

$$= \sum_{i=1}^{l} \alpha_i^0 y_i \mathbf{x}_i \cdot \mathbf{w}_0 \qquad \mathbf{w}_0 = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$$

$$= \sum_{i=1}^{l} \alpha_i^0 \left(1 - y_i b_0\right) \qquad y_i\left(\mathbf{x}_i \cdot \mathbf{w}_0 + b_0\right) = 1$$

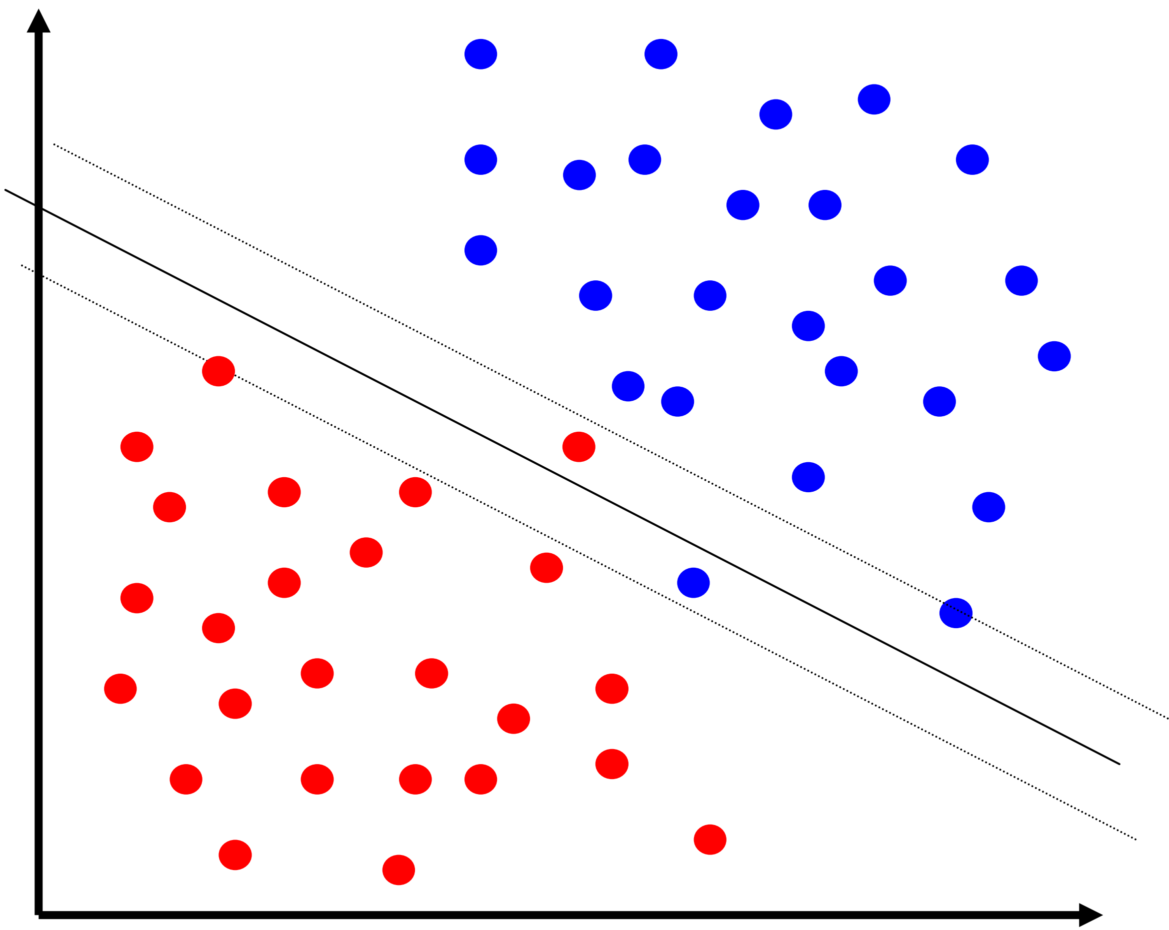
$$= \sum_{i=1}^{l} \alpha_i^0 \qquad \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$L(\mathbf{w}_0, b_0, \boldsymbol{\Lambda}_0) = \frac{1}{2} \mathbf{w}_0 \cdot \mathbf{w}_0$$

$\mathbf{w}_0$ の定める最適超平面のマージン ρ_0 は $2 \big/ \sqrt{\mathbf{w}_0 \cdot \mathbf{w}_0}$

$$L(\mathbf{w}_0, b_0, \boldsymbol{\Lambda}_0) = \frac{2}{\rho_0^2}$$

エラーを許す分類へ：ソフトマージン



線形分離が困難な場合、
各データごとに非負実数

$$\xi_i \geq 0 \quad \cdots \quad (1)$$

を用意し制約を緩和

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i$$
$$(i = 1, \ldots, l) \quad \cdots \quad (2)$$

マージンを広くし、誤差の和 $\sum_{i=1}^{l} \xi_i$ を
小さくする超平面で分離したい

$\frac{1}{2}\mathbf{w}^2 + C\sum_{i=1}^{l} \xi_i$ の最小化（C は定数）

ソフトマージン

$$\frac{1}{2}\mathbf{w}^2 + C\sum_{i=1}^{l}\xi_i \quad \cdots \quad (1)$$

を以下の制約のもと最小化

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l$$

$$\xi_i \geq 0, \quad i = 1, \ldots, l$$

$\mathbf{\Lambda}(\geq 0), \mathbf{R}(\geq 0)$ をラグランジュ乗数

$$L(\mathbf{w}, \boldsymbol{\xi}, b, \mathbf{\Lambda}, \mathbf{R}) =$$

$$\frac{1}{2}\mathbf{w}^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i[y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i] - \sum_{i=1}^{l} r_i\xi_i$$

(1) を最小化する $\mathbf{w}_0, b_0, \xi_i^0$ は以下の条件をみたす

$$\left.\frac{\partial L}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}_0} = \mathbf{w}_0 - \sum_{i=0}^{l}\alpha_i y_i \mathbf{x}_i = 0$$

$$\left.\frac{\partial L}{\partial b}\right|_{b=b_0} = \sum_{i=0}^{l}\alpha_i y_i = 0 \qquad \left.\frac{\partial L}{\partial \xi_i}\right|_{\xi_i=\xi_i^0} = C - \alpha_i - r_i = 0$$

$\mathbf{w}_0, b_0, \xi_i^0$ を $L(\mathbf{w}, \boldsymbol{\xi}, b, \mathbf{\Lambda}, \mathbf{R})$ に代入し変形

$$L(\mathbf{w}_0, \boldsymbol{\xi}^0, b_0, \mathbf{\Lambda}, \mathbf{R})$$

$$= \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

$L(\mathbf{w}_0, \boldsymbol{\xi}^0, b_0, \mathbf{\Lambda}, \mathbf{R})$ を最大化する $\mathbf{\Lambda}$ を

2 次計画法で計算

Kuhn - Tuckerの定理より

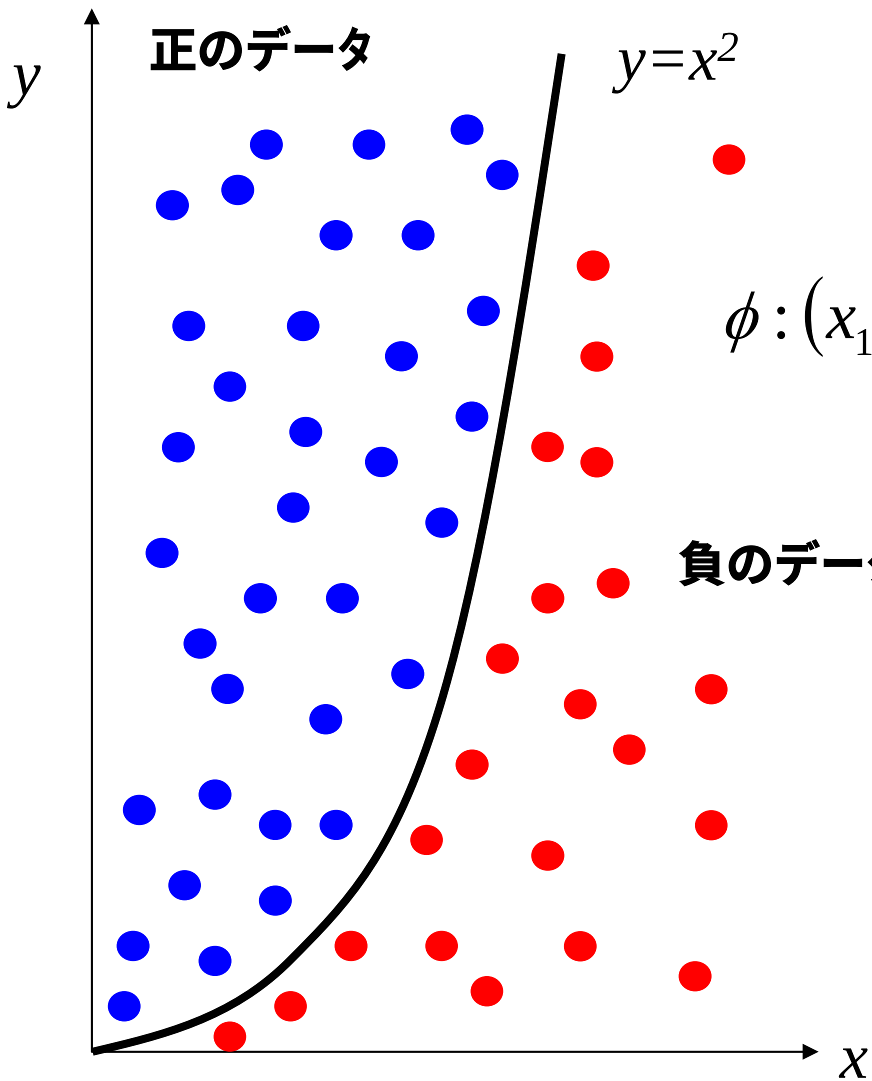
$$\alpha_i\left[y_i(\mathbf{x}_i \cdot \mathbf{w}_0 + b_0) - 1 + \xi_i^0\right] = 0$$

$\alpha_i \neq 0$ となる $\mathbf{x}_i$ をsupport vectorと定義.

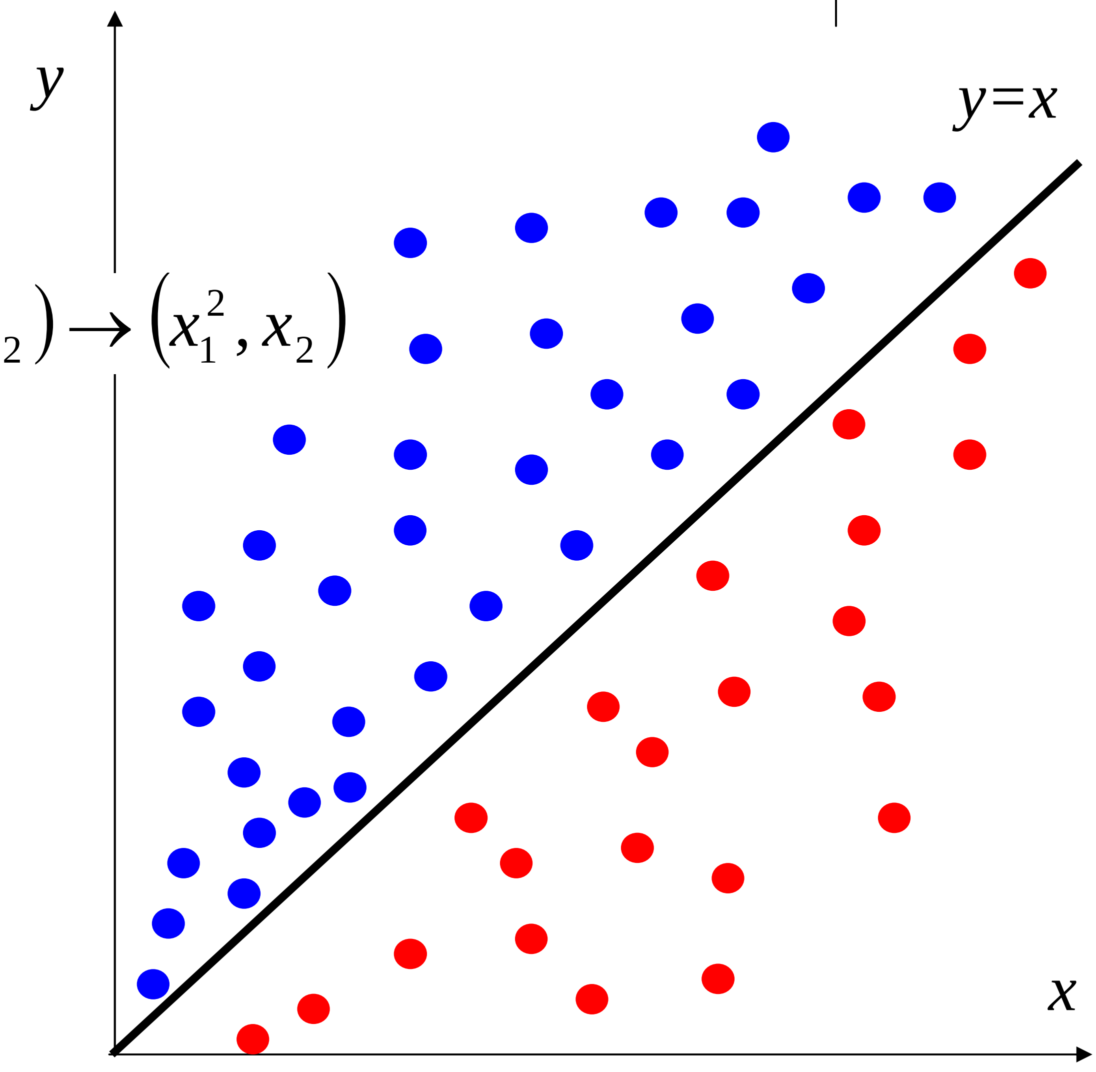
$\mathbf{w}_0$ はやはりsupport vector $\mathbf{x}_i$ の線形結合

$$\mathbf{w}_0 = \sum_{i=0}^{l}\alpha_i y_i \mathbf{x}_i$$

非線形の判別関数による分類



$$\phi:(x_1, x_2) \rightarrow (x_1^2, x_2)$$



非線形の判別関数： 特徴空間

入力空間 R^n の点を効果的に線形分　離できない場合、

より高い次元の特徴空　間 R^N に写像し線形分離を試　みる

$$\mathbf{x} \in R^n \to \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_N(\mathbf{x})) \in R^N$$

例　$\phi((x_1, x_2)) = \left(x_1^2, x_1 x_2, x_2^2, x_1, x_2\right)$

特徴空間で最適超平面　を形成する $\mathbf{w}_0$ は

$\phi(\mathbf{x}_i)$ の形をした support ve ctor の線形結合 $\sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i)$

非線形の判別関数： 計算量の軽減

入力空間に比べ巨大な特徴空間 $(n << N)$ を 使えば線形分離は容易に

例　訓練データ数より大きい次元の特徴空間.　$N \approx 2^n$ の場合も.

入力パラメータ数が $n = 100$ にたいして $N = 2^{100} \cong 10^{30}$ 次元の特徴空間

巨大な特徴空間では計算量も増大

双対問題を解くとき

$$L(\mathbf{w}_0, b_0, \boldsymbol{\Lambda}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \underline{\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)}$$

テスト $\mathbf{x}$ を最適超平面 $\mathbf{w}_0 = \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i)$ で分類するとき

$$\phi(\mathbf{x})\cdot\mathbf{w}_0 + b_0 = \sum_{i=1}^{l} \alpha_i y_i \, \underline{\phi(\mathbf{x})\cdot\phi(\mathbf{x}_i)} + b_0 \quad \geq 1? \quad (\leq -1?)$$

$\phi(\mathbf{x})$ と $\phi(\mathbf{y})$ を陽に計算せずに、$\phi(\mathbf{x})\cdot\phi(\mathbf{y})$ を計算可能な関数 ϕ の条件とは？

非線形の判別関数： カーネルトリック

多項式カーネル

$\mathbf{x},\mathbf{y}\in R^2$ のとき $\mathbf{x}=(x_1,x_2),\ \mathbf{y}=(y_1,y_2)$ と表現

$\phi(\mathbf{x})=(x_1^2,\sqrt{2}x_1x_2,x_2^2,\sqrt{2}x_1,\sqrt{2}x_2,1)\quad \phi(\mathbf{y})=(y_1^2,\sqrt{2}y_1y_2,y_2^2,\sqrt{2}y_1,\sqrt{2}y_2,1)$ ならば

$$\phi(\mathbf{x})\cdot\phi(\mathbf{y})=(\mathbf{x}\cdot\mathbf{y}+1)^2=\ (x_1y_1+x_2y_2+1)^2$$

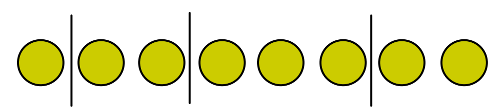
一般に $\mathbf{x},\mathbf{y}\in R^n$ のとき $\phi(\mathbf{x})\cdot\phi(\mathbf{y})=(\mathbf{x}\cdot\mathbf{y}+1)^d$ の形にするには...

$$(\mathbf{x}\cdot\mathbf{y}+1)^d=\sum_{c_1,c_2,\ldots,c_{n+1}}\frac{d!}{c_1!c_2!\cdots c_n!c_{n+1}!}(x_1y_1)^{c_1}(x_2y_2)^{c_2}\cdots(x_ny_n)^{c_n}1^{c_{n+1}}$$ ただし $\sum_{i=1}^{n+1}c_i=d,\ c_i\geq 0.$

$\phi(\mathbf{x})$ は $\sqrt{\frac{d!}{c_1!c_2!\cdots c_n!c_{n+1}!}}(x_1)^{c_1}(x_2)^{c_2}\cdots(x_n)^{c_n}1^{c_{n+1}}$ を各要素とするベクトルであればよい.

$\phi(\mathbf{x})$ の次元は ${}_{n+d}C_n={}_{n+d}C_d$

d=4, n=3



c_1=0, c_2=1, c_3=2, c_4=1

他の例：*Gauss* カーネル $\phi(\mathbf{x})\cdot\phi(\mathbf{y})=\exp\left(-\frac{|\mathbf{x}-\mathbf{y}|^2}{\sigma^2}\right)$

Support Vector Machine のまとめ

- 線形分離にマージンの概念を導入
- マージンを最適化する超平面を2次計画法で計算
  support vector によるマージン境界を表現
- エラー処理のためソフトマージンを導入
- 非線形な判別関数による分類　特徴空間の導入
- カーネルトリックによる計算の軽減