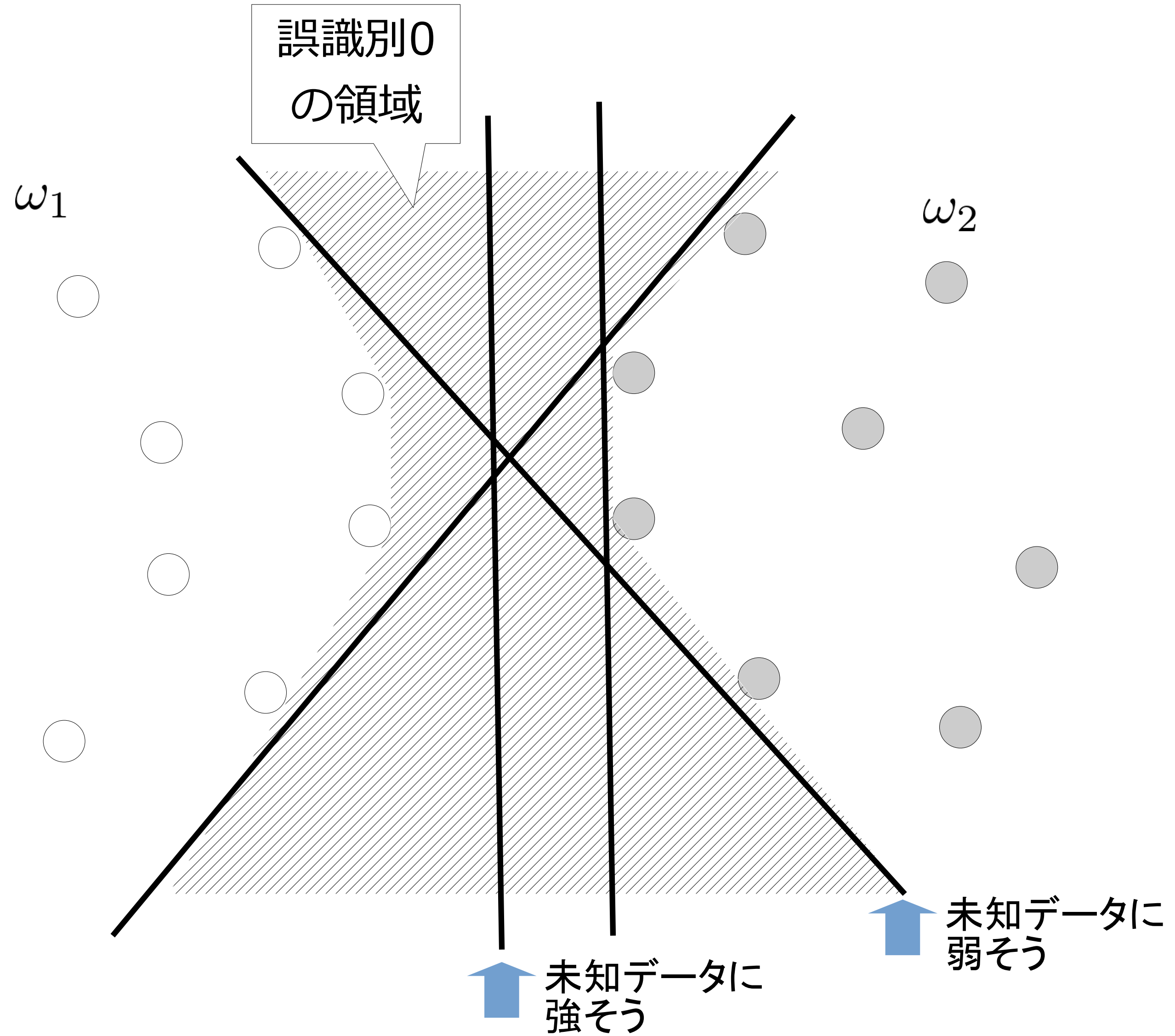
6. 限界は破れるか（1）
－　サポートベクトルマシン　－

- パーセプトロンの学習規則の限界
  - 学習パターンが線形分離可能である場合は識別面が見つかるが、信頼できる識別面とは限らない
  - 学習パターンが線形分離不可能である場合は、学習が停止しない

➡ サポートベクトルマシン（SVM）

6.1 識別面は見つかったけれど

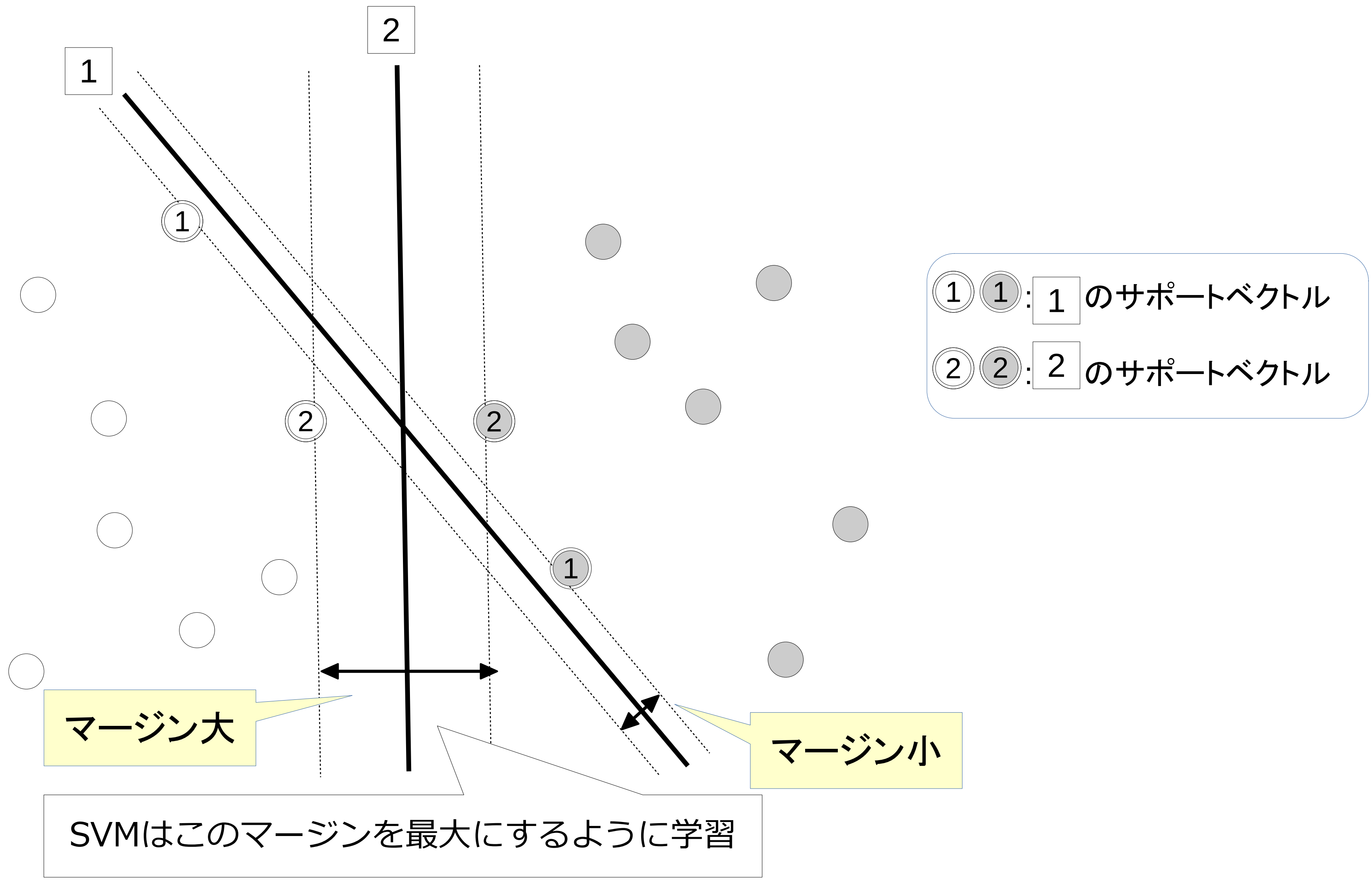
パーセプトロンの学習規則ではどれが見つかるかわからない



6.2 サポートベクトルマシンの学習アルゴリズム

6.2.1 サポートベクトル

- 線形SVM: マージン最大となる線形識別面を求める



6.2.2 マージンを最大にする

- 学習データ

$$\{(\boldsymbol{x}_i, y_i)\} \quad i = 1, \ldots, n, \quad y_i = 1 \ or \ -1$$

- 線形識別面の式

$$\boldsymbol{w}^T \boldsymbol{x} + w_0 = 0$$

- 識別面の制約の導入（係数を定数倍しても平面は不変）

$$\min_{i=1,\ldots,n} |\boldsymbol{w}^T \boldsymbol{x}_i + w_0| = 1$$

- 学習パターンと識別面との最小距離（＝マージン）

$$\min_{i=1,\ldots,n} Dist(\boldsymbol{x}_i) = \min_{i=1,\ldots,n} \frac{|\boldsymbol{w}^T \boldsymbol{x}_i + w_0|}{||\boldsymbol{w}||} = \frac{1}{||\boldsymbol{w}||}$$

これを
最大化

点と直線の距離の公式

$$r = \frac{|ax + by + c|}{\sqrt{a^2 + b^2}}$$

6.2.2 マージンを最大にする

- 目的関数の置き換え： $\min \frac{1}{2}||\boldsymbol{w}||^2$

極値が求めやすい２乗に

- 制約条件： $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1 \quad i = 1, \ldots, n$
- 解法：ラグランジュの未定乗数法
  - 問題 $\min f(x) \quad s.t.\ g(x) = 0$
  - ラグランジュ関数 $L(x, \alpha) = f(x) - \alpha g(x)$
    - $\alpha \geq 0$
    - x, α で偏微分して0になる値が極値

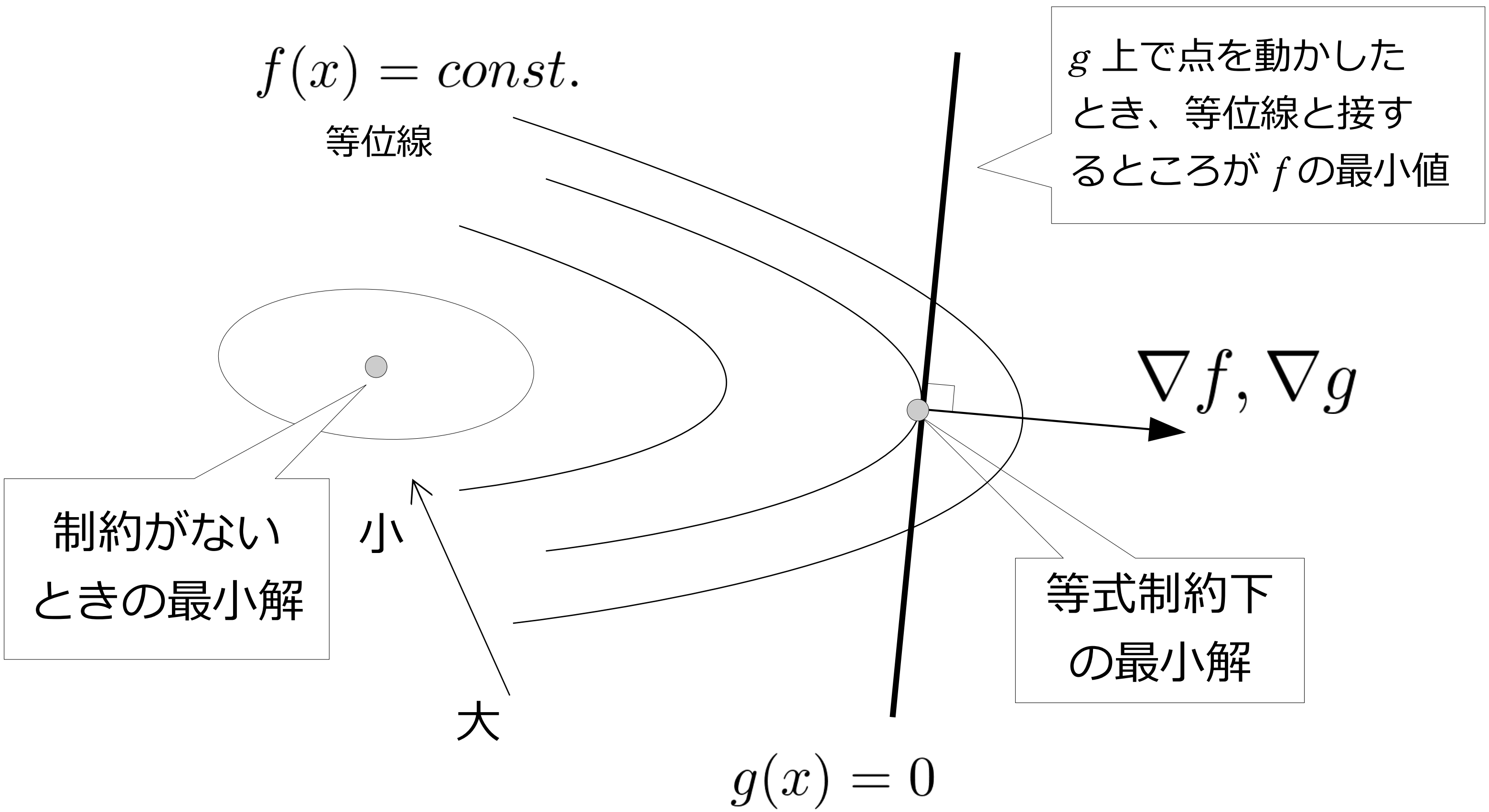
ラグランジュの未定乗数法（付録A.4）

$$\min f(x) \quad s.t.\ g(x) = 0$$

$$L(x, \alpha) = f(x) - \alpha g(x)$$

$$\frac{\partial L(\boldsymbol{x}, \alpha)}{\partial \boldsymbol{x}} = \nabla f(\boldsymbol{x}) - \alpha \nabla g(\boldsymbol{x}) = 0$$

$$\frac{\partial L(\boldsymbol{x}, \alpha)}{\partial \alpha} = -g(\boldsymbol{x}) = 0$$



6.2.2 マージンを最大にする

- 計算

$$L(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{n} \alpha_i (y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) - 1)$$

$$\frac{\partial L}{\partial w_0} = 0 \quad \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$$

$$\Downarrow$$

$$L(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j - \sum_{i=1}^{n} \alpha_i$$

$$\alpha_i \geq 0$$

最大化2次計画問題
→ $\boldsymbol{\alpha}$ が求まる

6.2.2 マージンを最大にする

- 定数項の計算
  - 各クラスのサポートベクトルから求める

$$w_0 = -\frac{1}{2}(\boldsymbol{w}^T \boldsymbol{x}_{s1} + \boldsymbol{w}^T \boldsymbol{x}_{s2})$$

- 識別関数

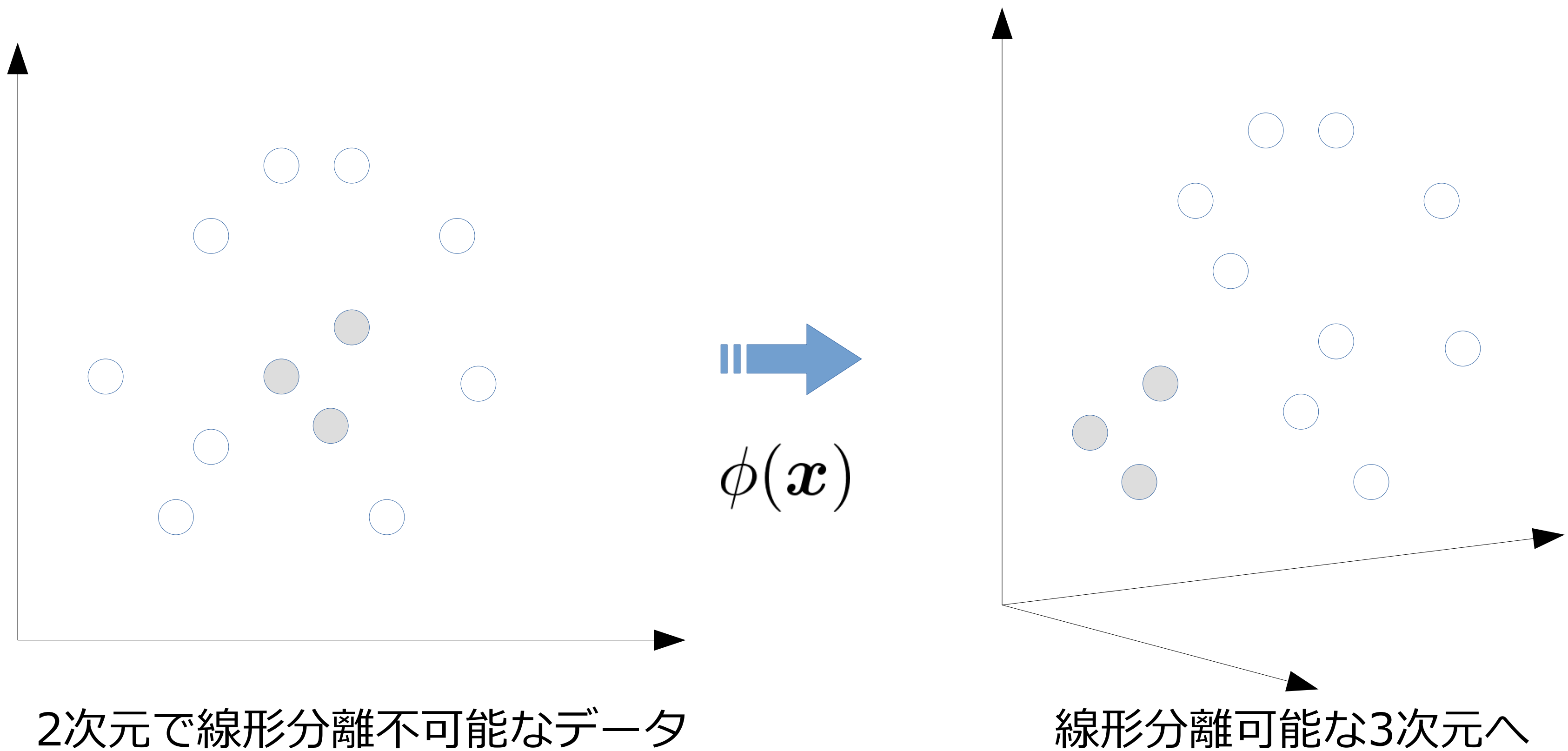
$$\begin{aligned} g(\boldsymbol{x}) &= \boldsymbol{w}^T \boldsymbol{x} + w_0 \\ &= \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}^T \boldsymbol{x}_i + w_0 \end{aligned}$$

サポートベクトルに対応するα_iのみが0以上、残りは0

6.3　線形分離可能にしてしまう

6.3.1　高次元空間への写像

- 特徴ベクトルの次元数を増やす



ただし、元の空間でのデータ間の
距離関係は保持するように

6.3.2　カーネル法

- 非線形変換関数： $\phi(\boldsymbol{x})$
- カーネル関数
  - 元の空間での距離が変換後の空間の内積に対応

$$K(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}')$$

  - カーネル関数の例
    - 多項式カーネル $K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T \boldsymbol{x}' + 1)^p$

$(\boldsymbol{x}^T \boldsymbol{x}')^p$ を用いる場合もある

    - ガウシアンカーネル $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{\sigma^2})$

これらの形であれば、対応する非線形変換が存在することが数学的に保証されている

6.3.2　カーネル法

- 変換後の識別関数：　$g(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + w_0$
- SVMで求めた $\boldsymbol{w}$ の値を代入

$$
\begin{aligned}
g(\boldsymbol{x}) &= \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}_i) + w_0 \\
&= \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + w_0
\end{aligned}
$$

非線形変換の
式は不要！！！

カーネルトリック

6.3.3　具体的なカーネル関数

- 多項式カーネル（2次、2次元）の展開

$$
\begin{aligned}
K(\boldsymbol{x}, \boldsymbol{x}') &= (\boldsymbol{x}^T \boldsymbol{x}')^2 \\
&= (x_1 x_1' + x_2 x_2')^2 \\
&= x_1^2 {x_1'}^2 + 2x_1 x_2 x_1' x_2' + {x_2}^2 {x_2'}^2 \\
&= (x_1^2, \sqrt{2} x_1 x_2, x_2^2) \cdot ({x_1'}^2, \sqrt{2} x_1' x_2', {x_2'}^2)
\end{aligned}
$$

$$
\begin{aligned}
K(\boldsymbol{x}, \boldsymbol{x}') &= (\boldsymbol{x}^T \boldsymbol{x}' + 1)^2 \\
&= (x_1 x_1' + x_2 x_2 + 1')^2 \\
&= x_1^2 {x_1'}^2 + {x_2}^2 {x_2'}^2 + 2x_1 x_2 x_1' x_2' + 2x_1 x_1' + \ 2x_2 x_2' + 1 \\
&= (x_1^2, x_2^2, \sqrt{2} x_1 x_2, \sqrt{2} x_1, \sqrt{2} x_2, 1) \cdot \\
&\qquad (x_1'^2, x_2'^2, \sqrt{2} x_1' x_2', \sqrt{2} x_1', \sqrt{2} x_2', 1)
\end{aligned}
$$

SVMの利点

- 学習は2次計画問題なので，必ず最適解が見つかる
- 求めるパラメータa_iの大半が0となるので，そのような状況に特化した高速な最適化アルゴリズム（たとえばSMO）を用いることができる
- カーネル関数を用いて，線形分離可能な高次元空間に特徴ベクトルを非線形写像することができる
  - 二つのデータ間にカーネル関数さえ定義できれば，元のデータが特徴ベクトルの形で表現されていなくてもよい