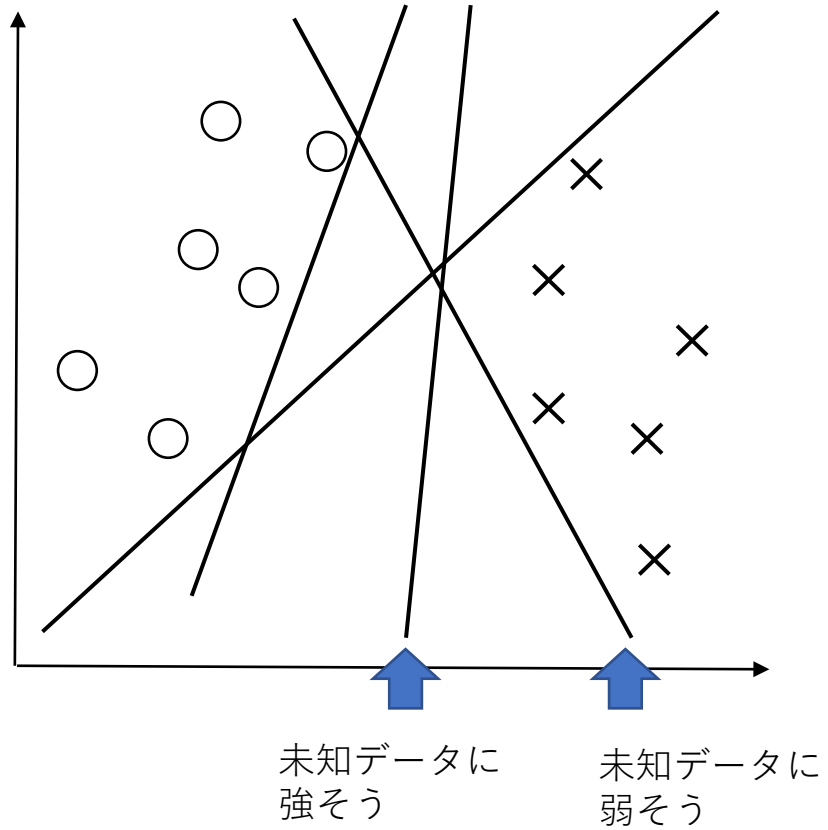
6. 限界は破れるか（1）－　サポートベクトルマシン　－

- パーセプトロンの学習規則の限界
  - 学習データ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
  - 学習データ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）

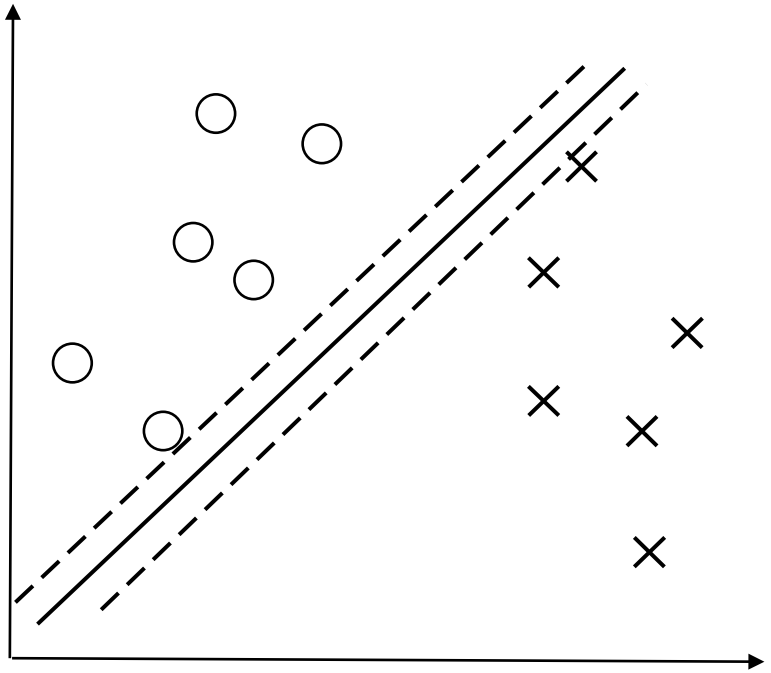
6.1　識別面は見つかったけれど

- 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

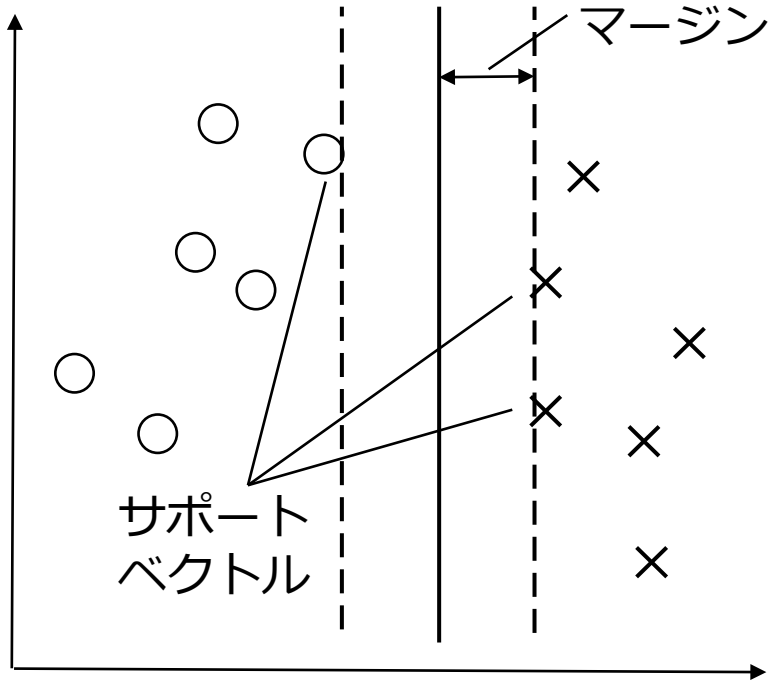


6.2 サポートベクトルマシンの学習アルゴリズム
6.2.1 サポートベクトル

- 線形SVM: マージン最大となる線形識別面を求める
  - マージン：識別面と最も近いデータとの距離（データは線形分離可能とする）



(a)マージンの小さい識別面



(b)マージンの大きい識別面

6.2.2 マージンを最大にする

- 学習データ

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

- 線形識別面の式

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

w, x はd次元

- 識別面の制約の導入（係数を定数倍しても平面は不変）

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

- 学習データと識別面との最小距離（＝マージン）

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

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：$\min \frac{1}{2}||\boldsymbol{w}||^2$
- 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
  - 例題（2変数、等式制約）　$\min f(x, y) \quad s.t.\ g(x, y) = 0$
  - ラグランジュ関数　$L(x, y, \alpha) = f(x, y) + \alpha g(x, y)$
    - ラグランジュ係数の制約 $\alpha \geqq 0$
    - x, α で偏微分して0になる値が極値

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} = \frac{\partial L}{\partial \alpha} = 0$$

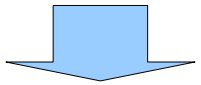
6.2.2 マージンを最大にする

- より解きやすい問題への変換

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

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

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



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

この式の最小化は $\boldsymbol{a}$ についての2次計画問題なので極値をとる $\boldsymbol{a}$ 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

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

- マージンが最大の識別関数
  - 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

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$$

$$= \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}^T\boldsymbol{x}_i + w_0$$

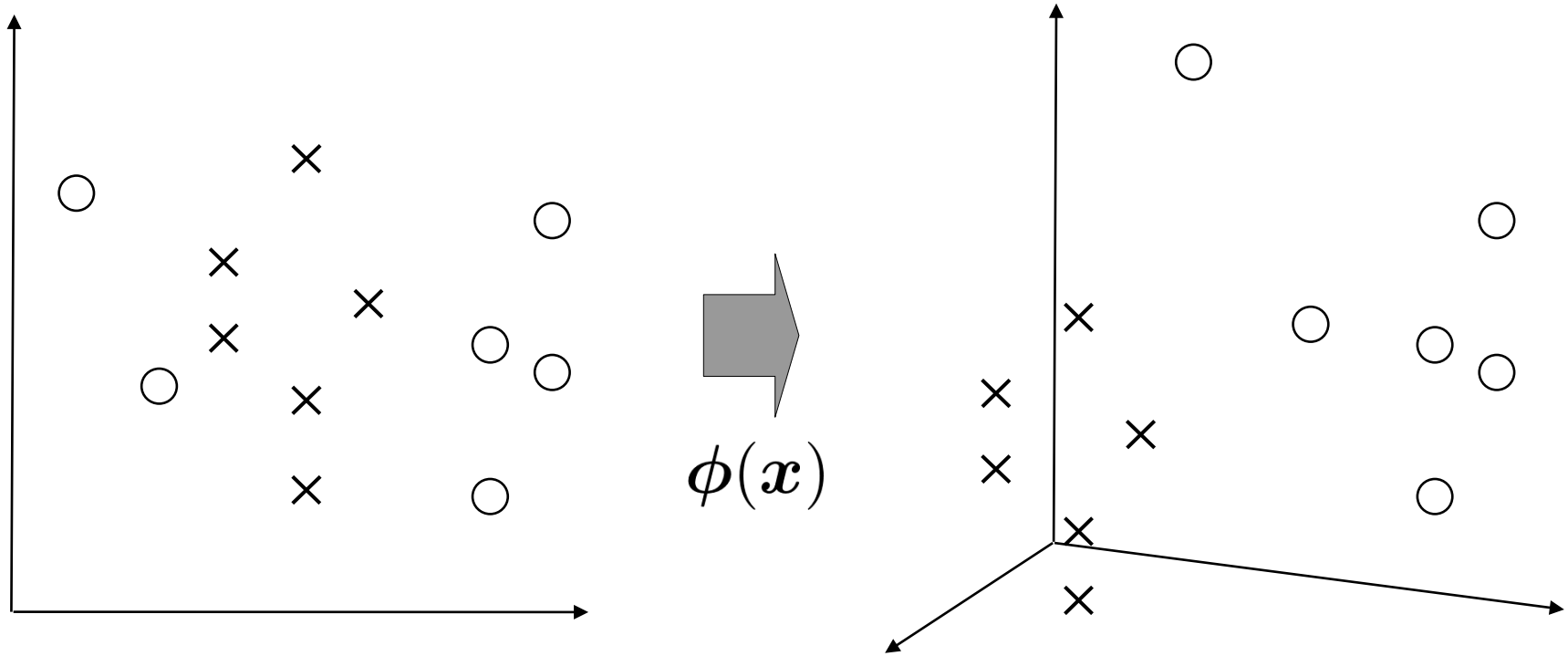
マージン最大の識別面の決定には
サポートベクトルしか関与しない

- 例題6.1

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

- クラスが複雑に入り交じった学習データ

  ⇒ 特徴ベクトルを高次元空間に写像



もとの次元で線形分離不可能なデータ

線形分離可能性の高い高次元へ

ただし、もとの空間でのデータ間の近接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：$\boldsymbol{\phi}(\boldsymbol{x})$
- カーネル関数

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

2つの引数値の近さを表す

  - 元の空間での距離が変換後の空間の内積に対応
  - カーネル関数の例
    - 多項式カーネル　$K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T \boldsymbol{x}' + r)^d$
    - ガウシアンカーネル　$K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma ||\boldsymbol{x} - \boldsymbol{x}'||^2)$
    - 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： $g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + w_0$
- SVMで求めた $\boldsymbol{w}$ の値を代入 $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i)$

$$g(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}_i) + w_0$$

$$= \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + w_0$$

非線形変換の式は不要！！！

カーネルトリック

  - 変換後の空間での線形識別面は、もとの空間での複雑な非線形識別面に対応

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
  - 6次元空間に写像されている

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

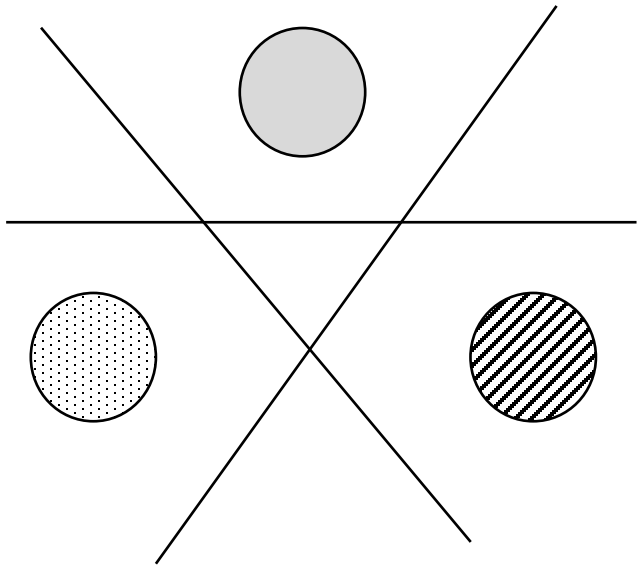
- 例題6.2

SVMの特徴

- 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α_iの大半が0となるので，この状況に特化した最適化アルゴリズム（たとえばSMO）で高速化が可能
- カーネル関数を用いて、特徴ベクトルを線形分離可能な高次元空間に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，元のデータがグラフのような特徴ベクトルの形で表現されていないものでもよい
- 2クラスの分類器なので、多クラスの分類には工夫が必要

2クラス分類器を用いた多クラス分類

- one-versus-rest法
  - 各クラスについて、そのクラスに属するかどうかを識別するSVMを作る
  - 2つ以上のクラスに属すると判定された場合は識別面からの距離が大きいものに分類する



2クラス分類器を用いた多クラス分類

- ペアワイズ法
  - クラス対ごとに識別器を作る
  - 判定は多数決を取る