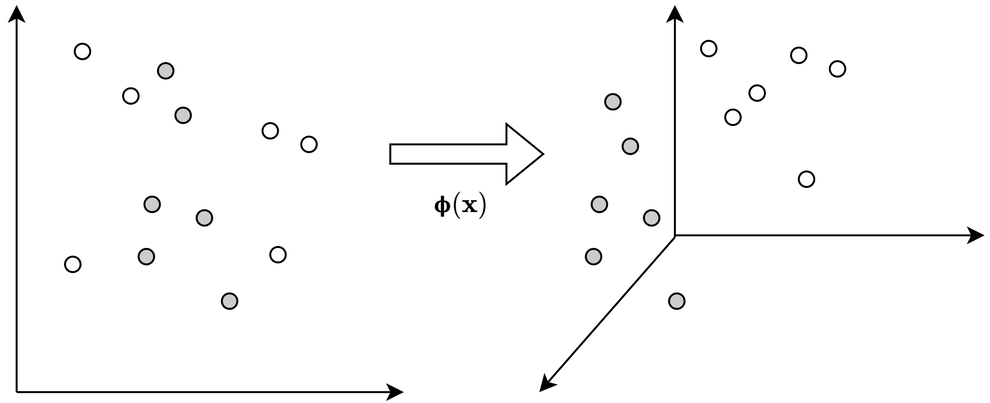
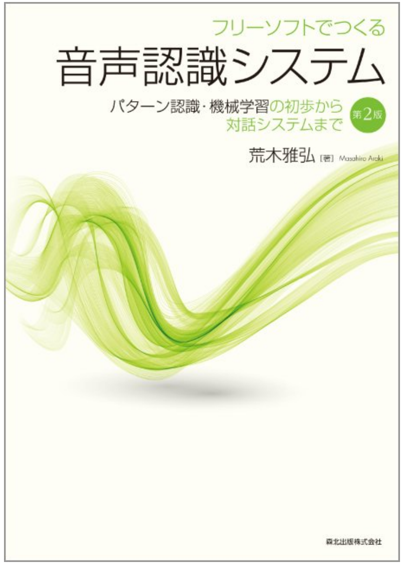
6. 限界は破れるか(1) － サポートベクトルマシン －



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

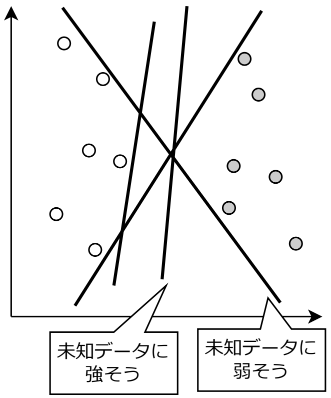


- 6.1 識別面は見つかったけれど
- 6.2 サポートベクトルマシンの学習アルゴリズム
- 6.3 線形分離可能にしてしまう

- 荒木雅弘:『フリーソフトでつくる音声認識システム(第2版)』(森北出版, 2017年)
- スライドとJupyter notebook
- サポートページ

6.1 識別面は見つかったけれど

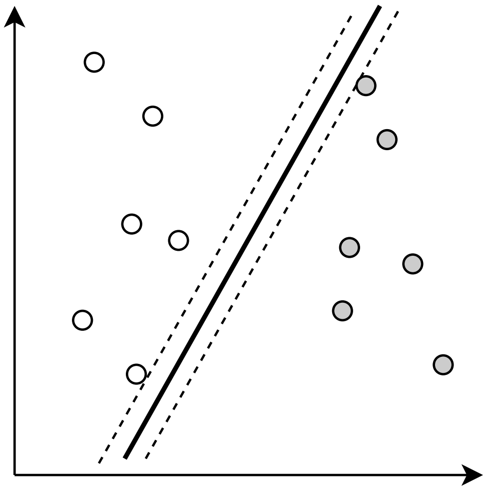
- パーセプトロンの学習規則の2つの限界
  1. 学習データが線形分離可能ならば識別面が見つかるが、信頼できる識別面とは限らない
  2. 学習データが線形分離不可能である場合は、学習が停止しない
- 限界1について(学習データは線形分離可能と仮定)
  - パーセプトロンの学習規則は全学習データが識別可能となった段階で停止するので、どの識別面が見つかるかわからない



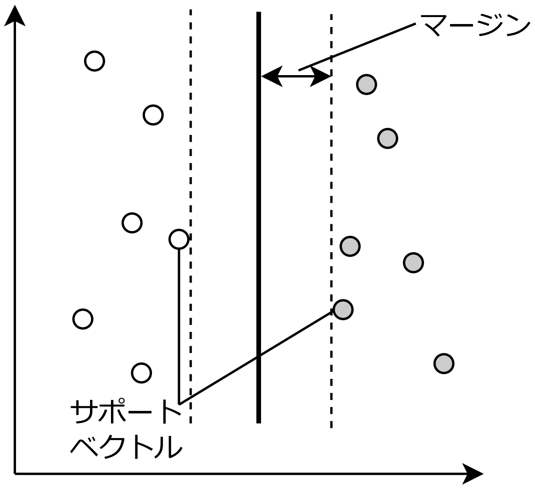
6.2 サポートベクトルマシンの学習アルゴリズム

6.2.1 サポートベクトル

- 線形サポートベクトルマシン(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(**サポートベクトル**)との距離



(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 (1/4)

- マージン最大化問題の定式化
  - 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 : $\boldsymbol{w}^T \boldsymbol{x} + w_0 = 0$
    - 係数に関する制約の導入(係数を定数倍しても平面は不変) : $\min_i |\boldsymbol{w}^T \boldsymbol{x}_i + w_0| = 1$
  - 最大化の対象 : データと識別面との距離 ($Dist$) の最小値(=マージン)

$$\min_i Dist(\boldsymbol{x}_i) = \min_i \frac{|\boldsymbol{w}^T \boldsymbol{x}_i + w_0|}{\|\boldsymbol{w}\|} = \frac{1}{\|\boldsymbol{w}\|}$$

- 参考)2次元平面での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（2/4）

- 目的関数の置き換え：$\min \frac{1}{2}\|\boldsymbol{w}\|^2$
  - 最大化対象の分母の最小化
  - 最小化問題として解きやすいように、2乗しておく
- 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）: $\min f(x, y) \;\; s.t. \; g(x, y) = 0$
    - 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 $\alpha \geq 0$
    - x, y, α で偏微分して0になる値が目的関数の極値

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

6.2.2 マージンを最大にする (3/4)

- より解きやすい問題への変換
  - $L(\boldsymbol{\alpha})$ の最小化は $\alpha_i \geq 0$ についての2次計画問題なので極値をとる $\boldsymbol{\alpha}$ が容易に求まる

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

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

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

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

6.2.2 マージンを最大にする (4/4)

- 定数項 w_0 は各クラスのサポートベクトルから求める

$$
\boldsymbol{w}^T \boldsymbol{x}_+ + w_0 = 1, \ \ \boldsymbol{w}^T \boldsymbol{x}_- + w_0 = -1
$$

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

- 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 のみが正の値、残りは 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

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

（補足）演習問題6.1

- 少数のデータが線形分離可能性を満たさない場合
  - i 番目のデータが制約を破っている度合いを表す変数 ξ_i (≥ 0) を導入し, 制約式を変更

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) \geq 1 - \xi_i \quad (i = 1, \ldots, n)$$

  - 目的関数に ξ_i の総和に重みを表す C を掛けて加える

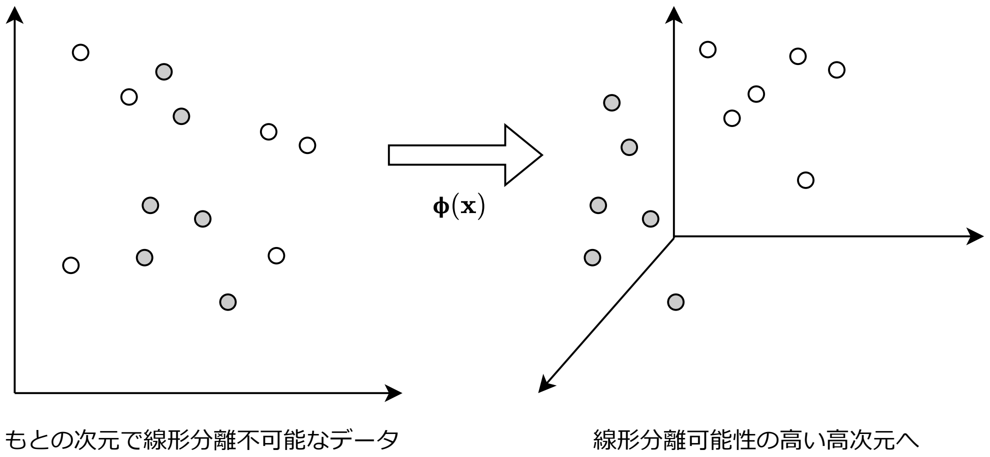
$$\min \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$

  - 解には $0 \leq \alpha_i \leq C$ という制約が加わるだけ

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

- 限界2の問題(線形分離不可能なデータ)への対処
  - クラスが複雑に入り交じった学習データを、線形分離可能性が高まる高次元空間に写像
  - ただし、もとの空間でのデータ間の近接関係は保持する



6.3.2 カーネル法 (1/2)

- 高次元空間への変換関数: $\boldsymbol{\phi}(\boldsymbol{x})$
- 2つのベクトル $\boldsymbol{x}, \boldsymbol{x}'$ の近さを表すカーネル関数 $K(\boldsymbol{x}, \boldsymbol{x}')$ を導入
    - 元の空間での近さを変換後の空間の内積に対応させる

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

    - カーネル関数の例
        - 多項式カーネル $K(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T \boldsymbol{x}' + r)^d$
        - ガウシアンカーネル $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$
        - これらのような、正定値性とよばれる性質を満たすカーネルであれば、対応する変換関数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 (2/2)

- 高次元空間での識別関数：$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + w_0$
- サポートベクトルマシンを適用し、$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{\phi}(\boldsymbol{x}_i)$ を代入

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

  - 変換関数が不要になった（＝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(特徴が2次元ベクトルで加算項が1, $d = 2, r = 1$)の展開

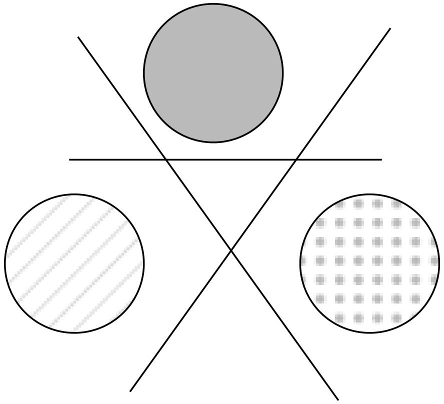
$$
\begin{aligned}
K(\boldsymbol{x}, \boldsymbol{x}') &= (\boldsymbol{x}^T \boldsymbol{x}' + 1)^2 \\
&= (x_1 x_1' + x_2 x_2' + 1)^2 \\
&= x_1^2 {x_1'}^2 + {x_2}^2 {x_2'}^2 + 2x_1 x_2 x_1' x_2' + 2x_1 x_1' + \ 2x_2 x_2' + 1 \\
&= (x_1^2, x_2^2, \sqrt{2}x_1 x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1) \cdot (x_1'^2,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

  - 6次元空間に写像されている

（補足）演習問題6.2

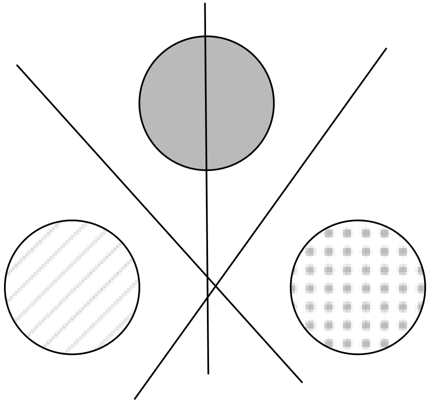
2クラス分類器を用いた多クラス分類（1/2）

- one-versus-rest法
  - 各クラスについて、そのクラスに属するか、他のクラスかを識別する 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



2クラス分類器を用いた多クラス分類 (2/2)

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



まとめ

- サポートベクトルマシンの特徴
  - 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性が高い高次元空間に非線形写像することができる
    - 2つのデータ間にカーネル関数さえ定義できれば, 元のデータはグラフや木構造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
- Jupyter notebook