One-Class Support Vector Machine OCSVM

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

One-Class SVM (OCSVM) とは？

✓サポートベクターマシン (Support Vector Machine, SVM) を領域推定問題に応用した手法

✓SVM では２つのクラス (1のクラス・-1のクラス) があったが、OCSVM では１クラスだけ (すべてのサンプルが同じクラス)

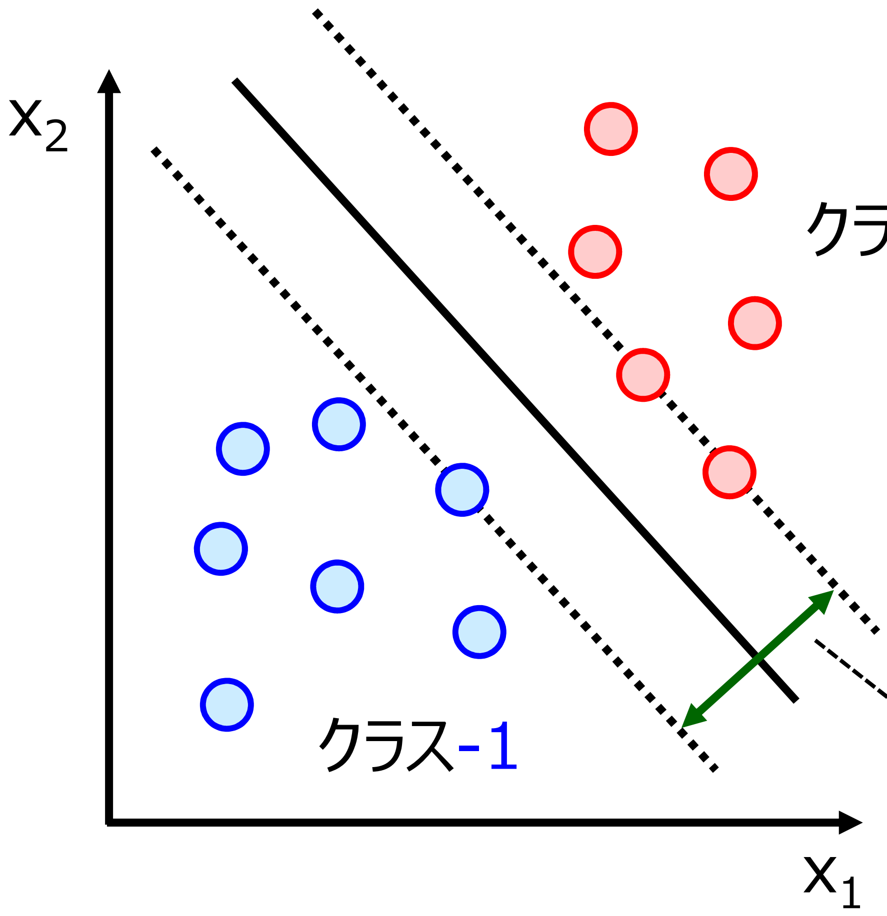
✓データ密度を連続的に推定できる

✓カーネルトリックにより非線形に拡張可能

✓外れ値検出・外れサンプル検出や、モデルの適用範囲の設定に応用される

- https://datachemeng.com/outlierdetection/
- https://datachemeng.com/numberofsamplesad/

SVMのおさらい：マージン最大化



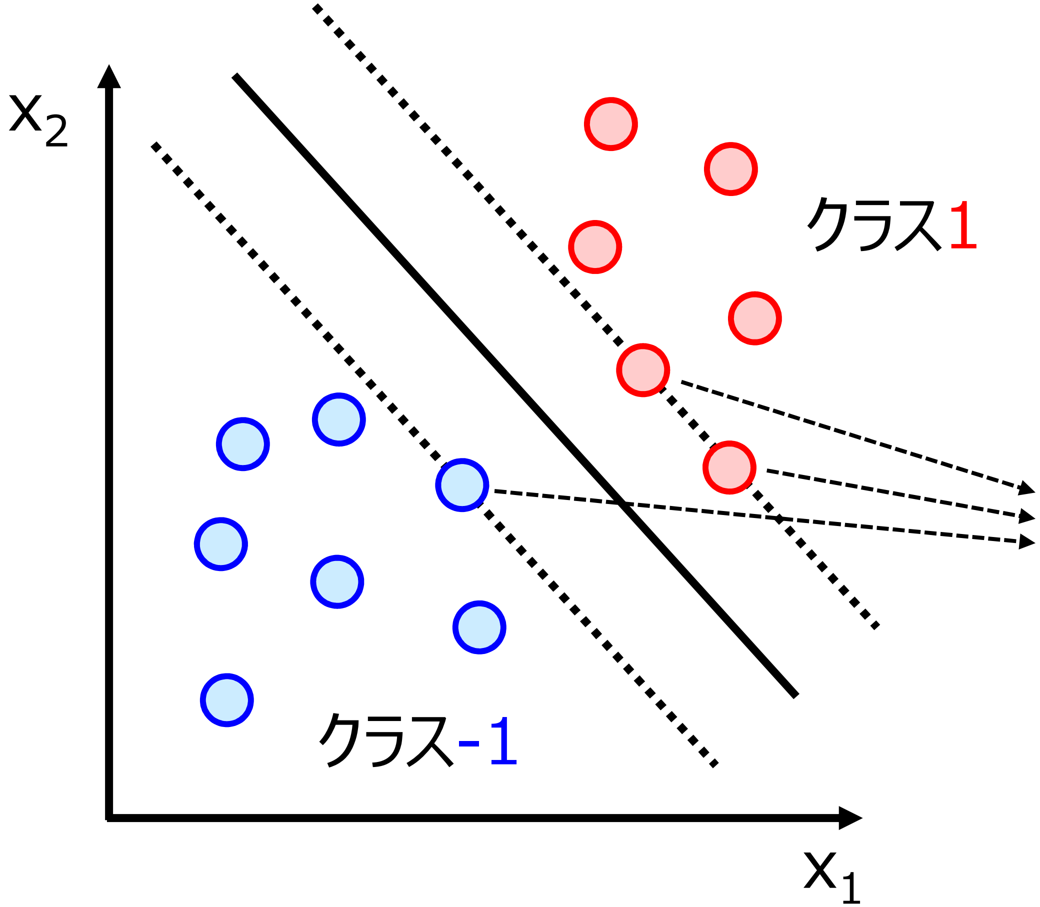
マージンを最大化するように
判別関数を決める！

$$f(x_1, x_2) = w_1 x_1 + w_2 x_2 + b$$
$$= \mathbf{x}\mathbf{w} + b$$

マージン $= \dfrac{2}{\|\mathbf{w}\|} = \dfrac{2}{\sqrt{w_1^{\ 2} + w_2^{\ 2}}}$

(点と直線との距離で計算)

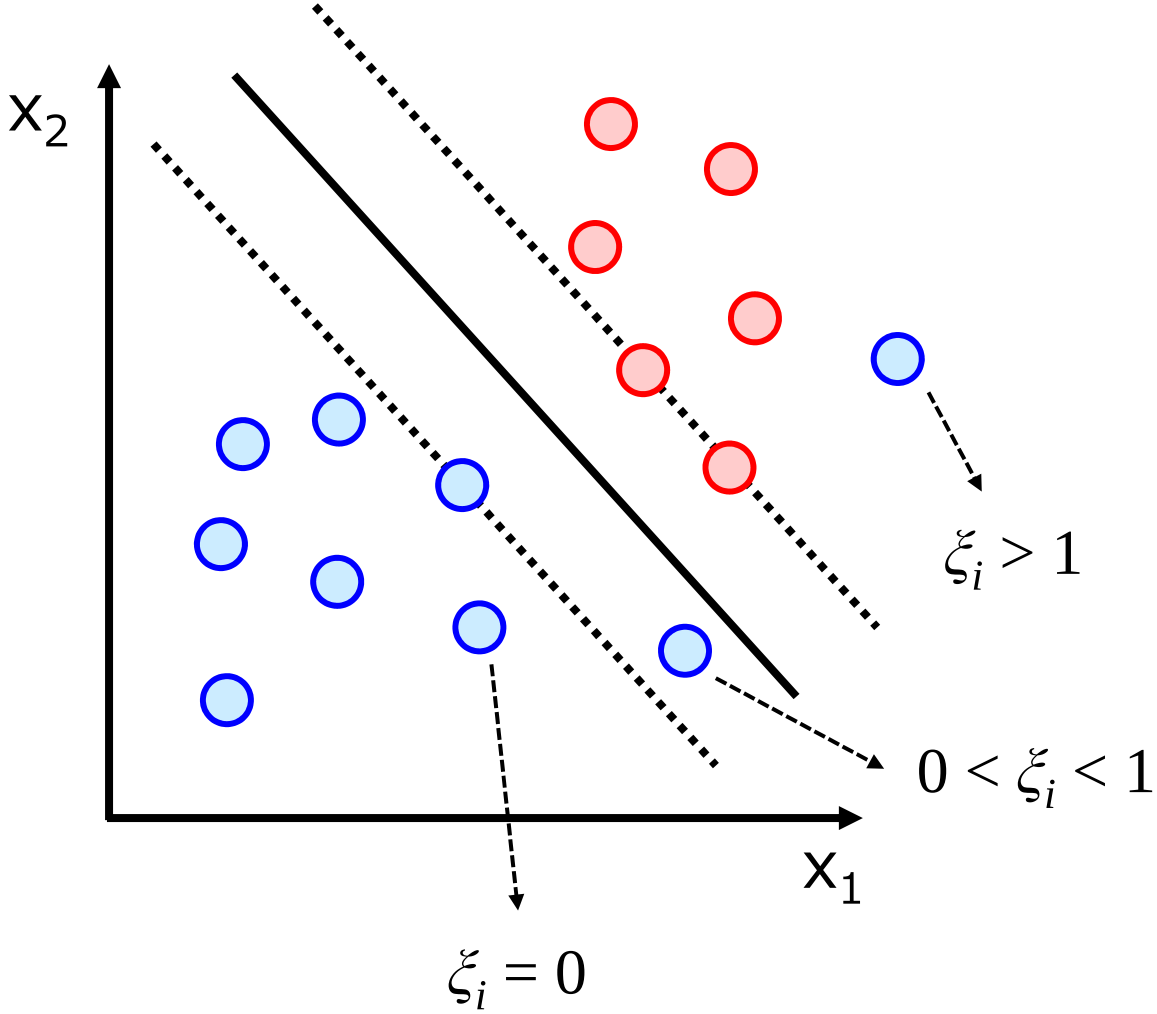
SVMのおさらい：サポートベクター



サポートベクター

・・・他のクラスのサンプルと
最も近いところにいるサンプル
$f(\mathbf{x}) = 1$ （クラス1）
$f(\mathbf{x}) = -1$ （クラス-1）

SVMのおさらい：スラック変数の導入



スラック変数 ξ を導入！

サンプルごとの ξ_i の和

$$\sum_{i=1}^{n} \xi_i$$

を最小化

n: モデル構築用サンプル数

SVMのおさらい：最小化する関数

✓ $\|\mathbf{w}\|$ / 2 の最小化 → 計算の都合上、$\|\mathbf{w}\|^2$ / 2 の最小化

✓ ξ_i の和の最小化

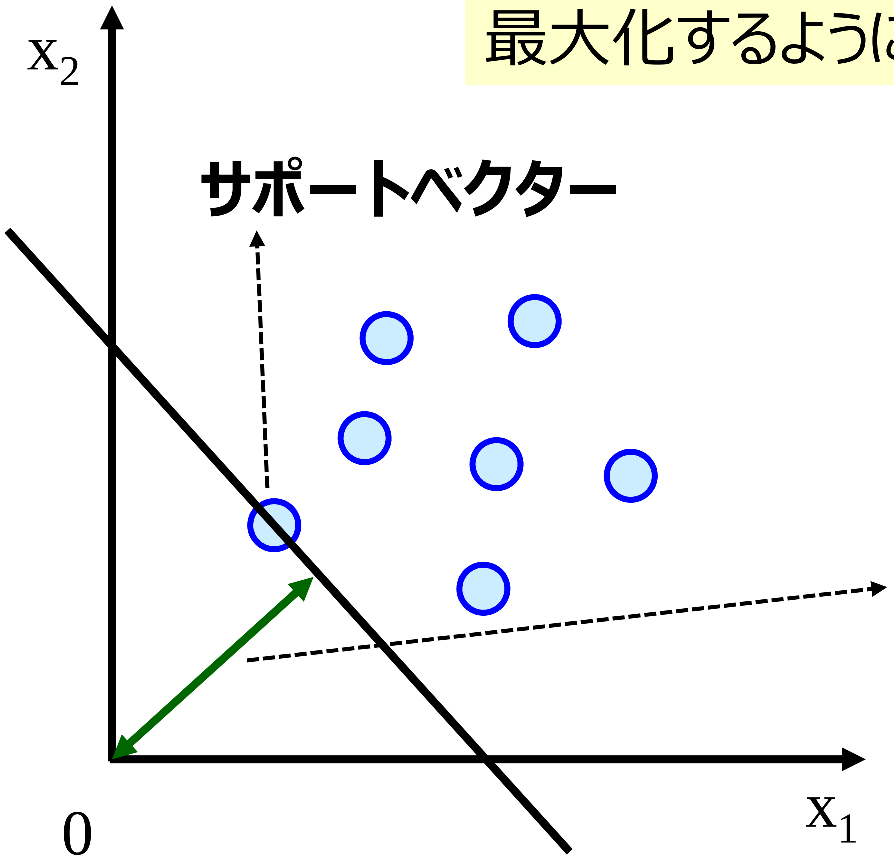
$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$ の最小化

ただし、 $\xi_i \geq 0, \quad y^{(i)} f\left(\mathbf{x}^{(i)}\right) \geq 1 - \xi_i$

C : ２つの項のバランスを決める係数
$\mathbf{x}^{(i)}$: i 番目のサンプルの説明変数
$y^{(i)}$: i 番目のサンプルの値 (1 もしくは -1)

OCSVMの考え方

原点からの距離 (**マージン**) を
最大化するように直線を決める！



$$f\left(x_1, x_2\right) = w_1 x_1 + w_2 x_2 - b$$
$$= \mathbf{x}\mathbf{w} - b \quad \left(b > 0\right)$$

マージン $= \dfrac{b}{\|\mathbf{w}\|} = \dfrac{b}{\sqrt{{w_1}^2 + {w_2}^2}}$

(原点と直線との距離で計算)

原点からの距離で大丈夫？

特に標準化 (オートスケーリング) したあととか、

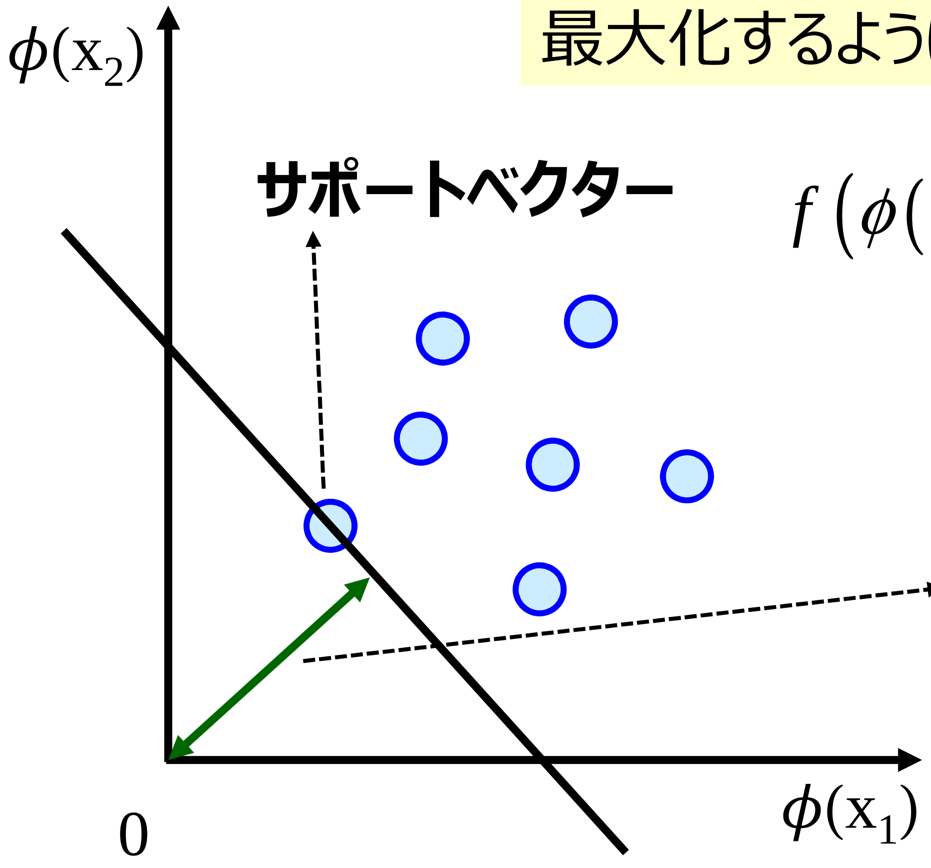
原点を中心にしてサンプルが散らばっているのでは？

外れ値ほど原点の近くになるように、

高次元空間に写像すれば大丈夫！

高次元空間で考える

原点からの距離 (**マージン**) を
最大化するように直線を決める！



$$f\left(\phi\left(x_1\right),\phi\left(x_2\right)\right)=w_1\phi\left(x_1\right)+w_2\phi\left(x_2\right)-b$$

$$=\phi\left(\mathbf{x}\right)\mathbf{w}-b \qquad \left(b>0\right)$$

マージン $= \dfrac{b}{\left\|\mathbf{w}\right\|} = \dfrac{b}{\sqrt{w_1{}^2+w_2{}^2}}$

(原点と直線との距離で計算)

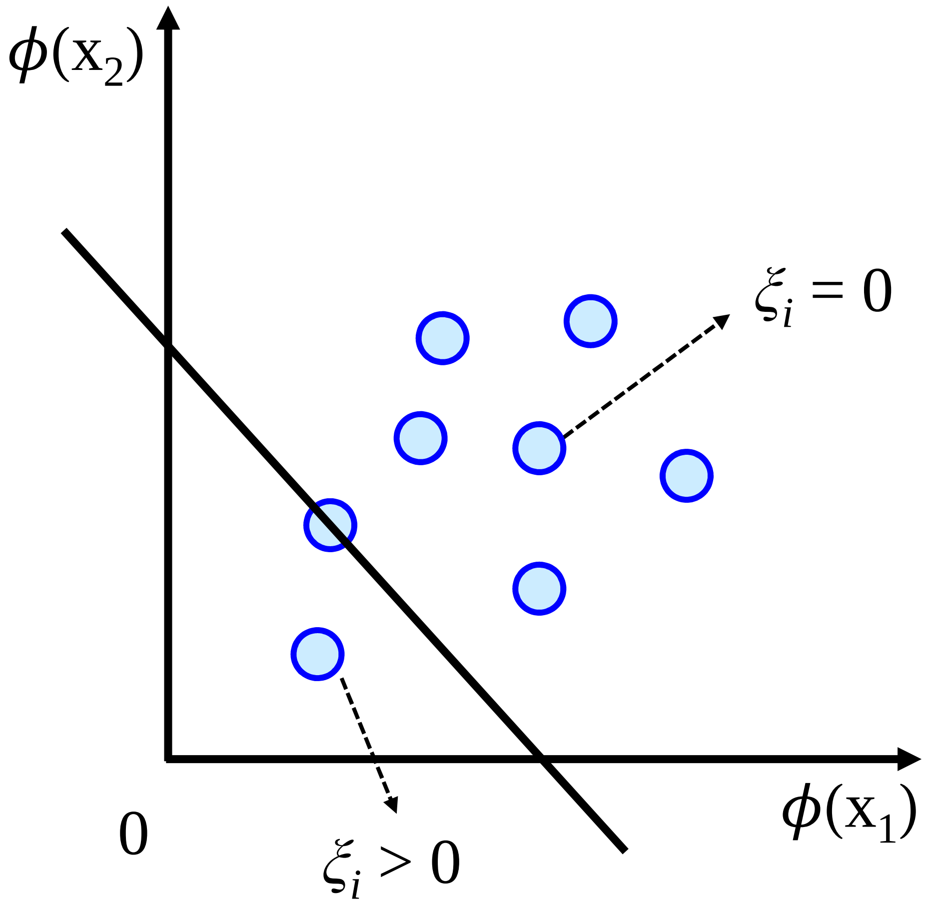
SVMと同様に、カーネルトリックにより ϕ そのものを考える必要はなくなる

マージンの最大化

$$\frac{b}{\|\mathbf{w}\|}$$ の最大化

⇒ $$\frac{1}{2}\|\mathbf{w}\|^2 - b$$ の最小化

原点付近にサンプルがあるときは？



スラック変数 ξ を導入！

サンプルごとの ξ_i の和

$$\sum_{i=1}^{n} \xi_i$$

を最小化

n: サンプル数

一緒に最小化

$$\frac{1}{2}\|\mathbf{w}\|^2 - b + \frac{1}{n\nu}\sum_{i=1}^{n}\xi_i$$

の最小化

重みを $\frac{1}{n\nu}$ という形でおいたのは、

あとで分かりやすくなるため

ν (ニュー) の範囲は、$0 < \nu \leq 1$

ただし、$\phi\left(\mathbf{x}^{(i)}\right)\mathbf{w} \geq b - \xi_i$

$$\xi_i \geq 0$$

$\mathbf{x}^{(i)}$: i 番目のサンプルの説明変数

重み w と b を求める

✓Lagrangeの未定乗数法

- ラグランジュ乗数α_i、β_i (i=1, 2, ..., n) を導入

$$G=\frac{1}{2}\|\mathbf{w}\|^2+\frac{1}{n\nu}\sum_{i=1}^{n}\xi_i-b-\sum_{i=1}^{n}\alpha_i\left\{\mathbf{w}\phi\left(\mathbf{x}^{(i)}\right)-b+\xi_i\right\}-\sum_{i=1}^{n}\beta_i\xi_i$$

$\mathbf{w}$、b、ξ_iに関してGを最小化し、α_i、β_iに関してGを最大化

⇩

$\mathbf{w}$、b、ξ_iに関してGが極小

⇩

G を$\mathbf{w}$、b、ξ_i それぞれで偏微分して 0 とする

偏微分して0

G を $\mathbf{w}$ で偏微分して0 $\quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi\left(\mathbf{x}^{(i)}\right)$

G を b で偏微分して0 $\quad \sum_{i=1}^{n} \alpha_i = 1$

G を ξ_i で偏微分して0 $\quad \alpha_i + \beta_i = \frac{1}{n\nu} \quad (i = 1, 2, \ldots, n)$

⇩

これらを使って
G を変形すると・・・

$$G = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \phi\left(\mathbf{x}^{(i)}\right) \phi\left(\mathbf{x}^{(j)}\right)^{\mathrm{T}}$$

また、 $\alpha_i + \beta_i = \frac{1}{n\nu}, \ \beta_i \geq 0$ より、 $0 \leq \alpha_i \leq \frac{1}{n\nu}$

二次計画問題

$$G = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\phi\left(\mathbf{x}^{(i)}\right)\phi\left(\mathbf{x}^{(j)}\right)^{\mathrm{T}}$$

制約　$\sum_{i=1}^{n}\alpha_i = 1$　　$0 \le \alpha_i \le \frac{1}{n\nu}$　のもとで、

G を α_i に対して最大化する二次計画問題を解くと α_i が求まる

⇨ $\mathbf{w}$ が求まる　　$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y^{(i)}\phi\left(\mathbf{x}^{(i)}\right)^{\mathrm{T}}$

直線の式を求める

$$f(\mathbf{x}) = \phi(\mathbf{x})\mathbf{w} - b$$

$$= \sum_{i=1}^{n} \alpha_i K\left(\mathbf{x}^{(i)}, \mathbf{x}\right) - b$$

K：カーネル関数　　$K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \phi\left(\mathbf{x}^{(i)}\right)\phi\left(\mathbf{x}^{(j)}\right)^{\mathrm{T}}$

カーネルトリックについてはこちら: https://datachemeng.com/supportvectormachine/

$$b = \frac{1}{n_{\mathrm{S}}} \sum_{j \in \mathrm{S}} \sum_{i=1}^{n} \alpha_i K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$$

S : サポートベクターの集合

n_{S} : サポートベクターの個数

G の変形

$$G = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\alpha_i - \alpha_i^{\ *}\right)\left(\alpha_j - \alpha_j^{\ *}\right)K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$$

$$-\varepsilon\sum_{i=1}^{n}\left(\alpha_i + \alpha_i^{\ *}\right) + \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^{\ *}\right)y_i$$

K：カーネル関数　　$K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \phi\left(\mathbf{x}^{(i)}\right)\phi\left(\mathbf{x}^{(j)}\right)^{\mathrm{T}}$

カーネルトリックについてはこちら

カーネル関数の例

✓線形カーネル

$$K\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right)=\mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}}$$

✓ガウシアンカーネル (使われることが多い)

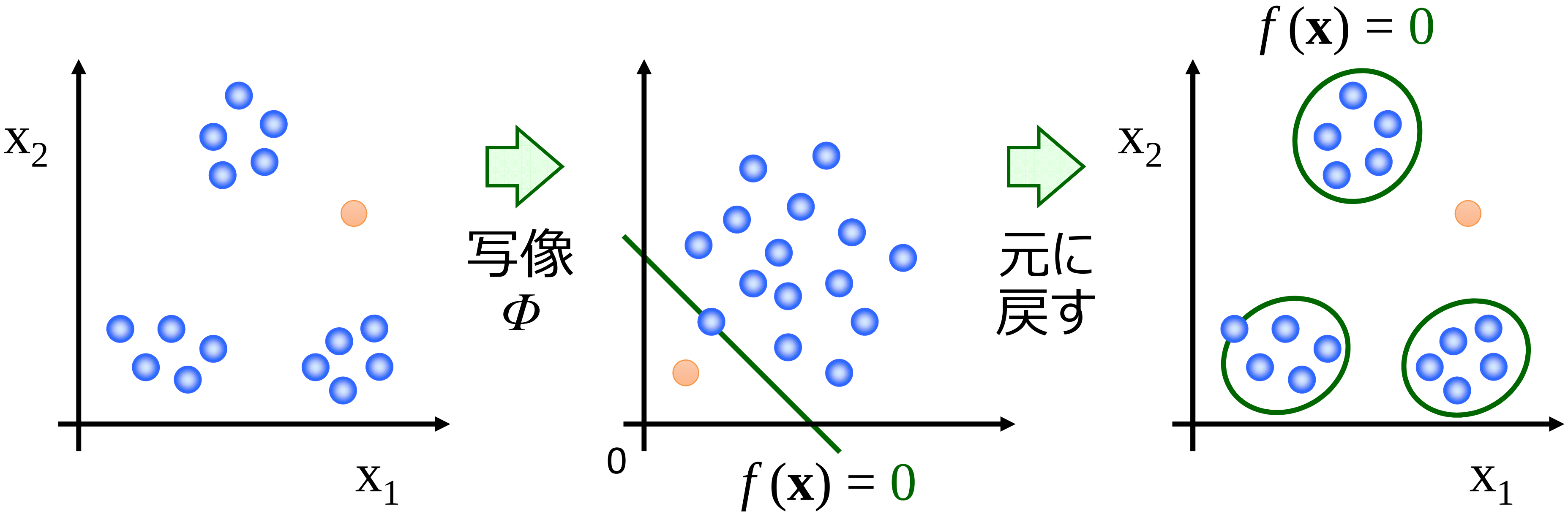
$$K\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right)=\exp\left(-\frac{\left\|\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right\|^2}{2\sigma^2}\right)=\exp\left(-\gamma\left\|\mathbf{x}^{(i)}-\mathbf{x}^{(j)}\right\|^2\right)$$

✓多項式カーネル

$$K\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right)=\left(1+\lambda\mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}}\right)^d$$

OCSVMのまとめ

✓Support Vector Machine (SVM) [1] を領域推定問題に応用



$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K\left(\mathbf{x}^{(i)}, \mathbf{x}\right) - b$$

カーネル関数 $K\left(\mathbf{x}^{(i)}, \mathbf{x}\right) = \phi\left(\mathbf{x}^{(i)}\right)^T \phi\left(\mathbf{x}^{(i)}\right)$

$$= \exp\left(-\gamma \left\| \mathbf{x}^{(i)} - \mathbf{x} \right\|^2\right)$$

[1] V.N. Vapnik, The nature of statistical learning theory, Springer, New York, 1999.

OCSVMの使い方

✓f($\mathbf{x}$) が 0 より小さくなったら、$\mathbf{x}$ は外れ値

✓f($\mathbf{x}$) の値が大きいほど、データ密度が高い

ν, γ の決め方

✓グラム行列 (すべてのサンプルの間でカーネル関数を計算した行列) の分散が最大になるように γ を最適化する

- グラム行列がばらついていたほうがデータセットをよく表現できるため
- γ の候補の例: $2^{-20}, 2^{-19}, \ldots, 2^{9}, 2^{10}$

✓$\nu = 0.003$ などとする

- ν は外れ値の割合の下限、といった意味合いをもつ
- ある変数が正規分布に従うとき、平均値から標準偏差の何倍かの範囲内にサンプルがある確率は決まる。1 からその確率を引くと、外れサンプルになる確率になる
  - (平均値 ± 標準偏差) 内にサンプルがある確率 : およそ 68.3 %
    $\rightarrow \nu = 1 - 0.683 = 0.317$
  - (平均値 ± 2×標準偏差) 内にサンプルがある確率 : およそ 95.5 %
    $\rightarrow \nu = 1 - 0.955 = 0.045$
  - (平均値 ± 3×標準偏差) 内にサンプルがある確率 : およそ 99.7%
    $\rightarrow \nu = 1 - 0.997 = 0.003$
- 結果が不適切であったときは、ν を変えて検討する