カーネル多変量解析

第4．4章　外れ値新規性検出

茨城大学工学部情報工学科
倉持辰洋

外れ値・新規性検出

- 非常に多いデータから意味のある重要な情報を抽出したい

例）・監視システムで通常と性質の異なるデータ

　　・数多くのニュースから新規性の高い情報を抽出

- この章では二つの方法を紹介する

１クラスν-サポートベクトルマシン
　⇒超平面で正常値と外れ値を分離

サポートベクトル領域記述法
　⇒球面で正常値と外れ値を分離

１クラス ν-サポートベクトルマシン（１）

- νトリックを使った外れ値検出
- 特徴ベクトルωとパラメータ Φ の内積からなる関数

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x})$$

を考える

- サンプルは $f(\boldsymbol{x}^{(1)}), \ldots, f(\boldsymbol{x}^{(n)})$ のように1次元のデータとなる

１クラス ν- サポートベクトルマシン（２）

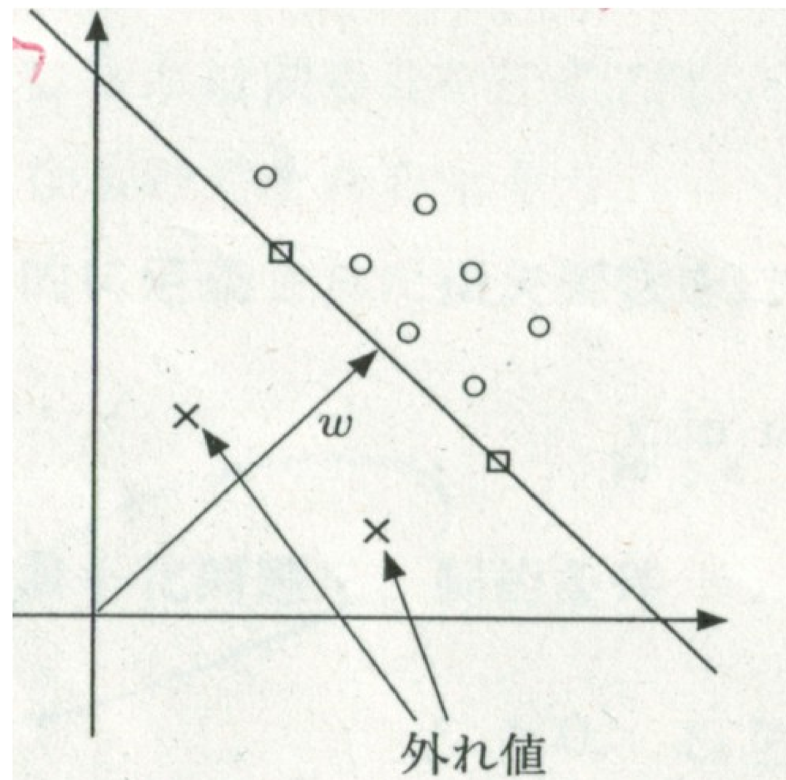
- しきい値ρ（＞０）でデータを二つに分ける

$\rho \le f\left(\boldsymbol{x}^{(i)}\right)$ となるデータ

⇒正常値

$\rho > f\left(\boldsymbol{x}^{(i)}\right)$ となるデータ

⇒外れ値



しきい値ρ(1)

- ρの値を決めたい

  正常値のクラスは多くのデータを含むはず

  　⇒　ρの値は小さいほうがいい

  クラスタはできるだけまとまってほしい

  　⇒　ρの値は大きいほうがいい

- ４．３章の関数（４．４２）を使い損失関数

$$r_\rho(f(\boldsymbol{x})) = max\{0, \rho - f(\boldsymbol{x})\}$$

を作り、外れ値で正の値をとるようにする。

しきい値ρ(2)

- 二次の正則化を行う

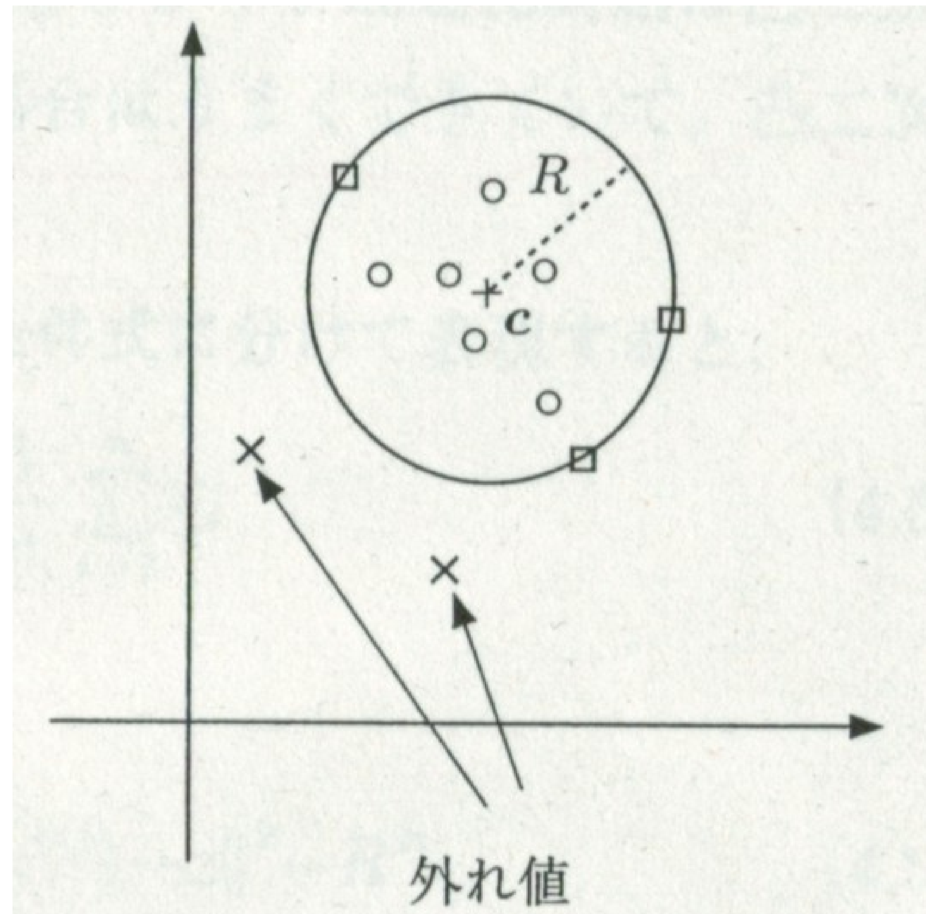
$$\min_{\rho>0} \frac{1}{n}\sum_{i=1}^{n} r_\rho\left(f\left(\boldsymbol{x}^{(i)}\right)\right) + \frac{1}{2}\boldsymbol{\alpha}^T K \boldsymbol{\alpha} - \nu\rho$$

- この式はが入っていない-サポートベクトルマシンと同じものとなっている。

　⇒正常値であるサンプルを一つのクラスとして扱う1クラス問題として解ける

サポートベクトル領域記述法(1)

- 特徴空間において
  中心：**c**
  半径：R
  というパラメータをとる
- 中心からの二乗距離が
  球からはみ出す場合
  損失とする。



サポートベクトル領域記述法(2)

- 球からはみだす、つまりR^2以上のときを損失として加えてやるので

$$\min_{R^2, \boldsymbol{c}} \frac{1}{n} \sum_{i=1}^{n} r_{R^2}\left(\left\|\boldsymbol{\Phi}\left(\boldsymbol{x}^{(i)}\right)-\boldsymbol{c}\right\|^2\right)+\nu R^2$$

となり、損失関数は

$$r_{R^2}(z)=max\{0, z-R^2\}$$

となる

サポートベクトル領域記述法(3)

- 正則化項 νR^2 ⇒半径が大きくなるのを抑止
- r_{R^2} ⇒区分的な二つの制約式に分けれる

↓けれど

- 制約式にパラメータcの２次式が入っている

  ⇒一般に、制約が2次関数である最適問題は凸二次計画問題になるとは限らない

- 今回の場合は凸二次計画問題となる

凸二次計画問題へ(1)

- 損失関数を二つの条件式に分けて表現

  最小化問題:

$$\min_{R^2,\xi,\boldsymbol{c}} vR^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i$$

  制約条件:　$\xi_i \geq 0,\quad \xi_i \geq \|\boldsymbol{\Phi}\left(\boldsymbol{x}^{(i)}\right) - \boldsymbol{c}\|^2 - R^2$

  これらと等価になる

凸二次計画問題へ(2)

- 先の最小化問題より
  ラグランジュ関数:

$$L\left(R^2, \boldsymbol{\xi}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right) = vR^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\beta_i\xi_i - \sum_{i=1}^{n}\gamma_i\left(\xi_i - \left\|\boldsymbol{\Phi}\left(\boldsymbol{x}^{(i)}\right) - \boldsymbol{c}\right\|^2 + R^2\right)$$

- １次式のみの R^2, ξ_i について係数を0とおくと

$$v = \sum_{i=1}^{n}\gamma_i, \qquad \frac{1}{n} - \beta_i = \gamma_i$$

- $\boldsymbol{c}$ について微分し0とおくと

$$\boldsymbol{c} = \frac{1}{\sum_{i=1}^{n}\gamma_i}\sum_{i=1}^{n}\gamma_i\boldsymbol{\Phi}\left(\boldsymbol{x}^{(i)}\right) = \frac{1}{v}\sum_{i=1}^{n}\gamma_i\boldsymbol{\Phi}\left(\boldsymbol{x}^{(i)}\right)$$

凸二次計画問題へ（3）

- $v=\sum_{i=1}^{n}\gamma_i$ をLに代入することで双対問題の目的関数が得られる

$$L_{dual}(\boldsymbol{\gamma})=\sum_{i=1}^{n}\gamma_i\left\|\boldsymbol{\Phi}\left(\boldsymbol{x}^{(i)}\right)-\frac{1}{v}\sum_{j=1}\gamma_i\boldsymbol{\Phi}\left(\boldsymbol{x}^{(j)}\right)\right\|^2$$

$$\sum_{i=1}^{n}\gamma_i\left\{K_{ii}-\frac{2}{v}\sum_{j=1}^{n}\gamma_j K_{ij}+\frac{1}{v^2}\sum_{j,j'}\gamma_j\gamma_j{}'K_{jj}{}'\right\}$$

$$\sum_{i=i}^{n}\gamma_i K_{ii}-\frac{1}{v}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_i\gamma_j K_{ij}$$

- 凸二次計画問題となった

凸二次計画問題へ（4）

- 制約式：

$$0 \leq \gamma_i \leq \frac{1}{n}, \qquad \sum_{i=1}^{n} \gamma_i = v$$

パラメータRを求める

- 相補性条件より

  γ_i, β_i が0でない場合

  $\xi_i = 0, \xi_i - \|\boldsymbol{\Phi}(\boldsymbol{x}^{(i)}) - \boldsymbol{c}\|^2 + r^2 = 0$

- これより以下の式で求めることができる

$$R^2 = \|\boldsymbol{\Phi}(\boldsymbol{x}^{(i)}) - \boldsymbol{c}\|^2$$