# カーネル多変量解析
# 第4 章凸計画問題を用いたカーネル多変量解析（３）

サ ミンソン

# 4.4 外れ値・新規性検出

- 監視システムで通常の観測と性質の異なるデータが得られたときにそれを検知、ネット上の配信されるニュースで新規性の高い情報を取り出す

＝多くの情報から重要な情報を抽出

→データマイニングの原点とも言える問題

# (a)1クラス $\nu$ サポートベクトルマシン

- 特殊ベクトルとパラメータの内積からなる関数

  $f(x) = \omega^T \varnothing(x)$

- サンプル : $f(x^{(1)}), \cdots, f(x^{(n)})$
- このデータをしきい値 $p > 0$ で二つに分け、

  $p \leq f(x^{(i)})$ : 正常値

  $p > f(x^{(j)})$ : 外れ値

# つづき

- 正常値のクラスはたくさんのデータを含むので、pは小さいほどよい。
- 一方、クラスタはできるだけまとまっていてほしい→pはできるだけ大きいほうがよい
- 関数4.42を使って

$$r_p(f(x)) = \max\{0, p - f(x)\}$$

という損失関数を作り、外れ値では正の値を取る。この損失を抑えながらpを大きくする

# つづき

- リプレゼンター定理のために2次の正則化を行うと、

$$\min_{p>0} \frac{1}{n}\sum_{i=1}^{n} r_p(f(x^{(i)})) + \frac{1}{2}\alpha^T K a - \nu p$$

- という最適化問題を解くことに帰着され、これは $y^{(i)}$ が入っていない $\nu$ −サポートベクトルマシン(4.44)と同じもの
- 外れ値でないサンプルを一つのクラスとして扱う1クラス問題とみなすことができるので1クラス $\nu$−サポートベクトルマシンと呼ばれる

# (b)データを包含する球

- 球面で分難する・・・サポートベクトル領域記述法(SVDD)
- できるだけ小さな半球の球にたくさんのサンプルが入るような規準を立てる

# つづき

- 特殊空間における中心c、半径Rのパラメータを取り、中心からの二乗距離が球からはみ出る場合、$R^2$ 以上になった部分を損失として加える

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

- 損失関数

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

つづき

つづき

# つづき

$$L_{dual(\gamma)} = \sum_{i=1}^{n} \gamma_i \| \varnothing(x^{(i)}) - \frac{1}{\nu} \sum_{j=1}^{n} \gamma_j \varnothing(x^{(j)}) \|^2$$

$$\min_{R^2, \varepsilon, c} \nu R^2 + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \qquad \varepsilon_i \geq 0, \varepsilon_i \geq \| \varnothing(x^{(i)}) - c \|^2 - R^2$$

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

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

$$\frac{1}{n} - \beta_i = \gamma_i \qquad \gamma_i \neq 0, \beta_i \neq 0$$

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

$$c = \frac{1}{\sum_{i=1}^{n} \gamma_i} \sum_{i=1}^{n} \gamma_i \varnothing(x^{(i)}) = \frac{1}{\nu} \sum_{i=1}^{n} \gamma_i \varnothing(x^{(i)}) \qquad \nu = \sum_{i=1}^{n} \gamma_i$$

$$L(R^2, \varepsilon, c, \beta, \gamma) = \nu R^2 + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i - \sum_{i=1}^{n} \beta_i \varepsilon_i - \sum_{i=1}^{n} \gamma_i (\varepsilon_i - \| \varnothing(x^{(i)}) - c \|^2 + R^2)$$

$$\min_{R^2,c} \frac{1}{n} \sum_{i=1}^{n} r_{R^2}(\| \varnothing(x^{(i)}) - c \|^2) + vR^2$$