7.1.4 回帰のためのSVM

7.1.3項までは回帰問題に対するSVMを考えてきた。SVMの利点は、正しく予測できている訓練データについては予測の際に無視して良い、つまりカーネル法における疎な解が得られることであった。

回帰のためのSVMでも同様に疎な解のカーネル法が得られる。すなわち、正しく回帰予測できている訓練データについては予測の際に用いず、回帰線から離れているデータのみに基づいて予測するモデルが得られる。

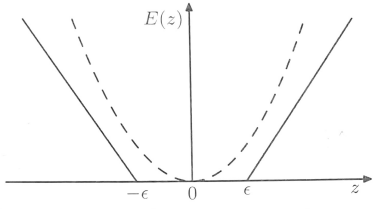
本節では回帰のためのSVMを定式化し、予測のためのモデルを導くところまで行う。

まず、「正しく回帰予測できている」ということを表現するために、パラメータ ε を導入する。予測値 $y(x)$ と観測値 t の差が $\varepsilon\ (>0)$ 未満の場合には誤差が0であると考え、それ以外の場合には線形に誤差が増加していくと考えると、次のような ε 許容誤差関数を構築できる。

これを2とする方が都合が良い.

$$E_\varepsilon(z) = E_\varepsilon(t - y(x))$$

$$= \begin{cases} 0 & \text{if } |y(x)-t| < \varepsilon \\ |t - y(x)| - \varepsilon & \text{otherwise} \end{cases}$$



この誤差関数を、これまで扱ってきた回帰の誤差関数に代用することで、SVMの誤差関数の原型が得られる。3章で扱った回帰モデルは二乗和誤差を使った正則化付き誤差関数を用いた。

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - y_n\}^2 + \frac{\lambda}{2}\|w\|^2$$

二乗和の代わりに ε 許容誤差関数を代用し、定数倍すると次の誤差関数が得られる。

$$C\sum_{n=1}^{N} E_\varepsilon(t_n - y(x_n)) + \frac{1}{2}\|w\|^2 \qquad (7.52)$$

SVMでは、この誤差関数を最小化することを目的とするが、絶対値があるままだと、最適化問題として解くことが難しい。絶対値を避けるためにスラック変数を導入する。

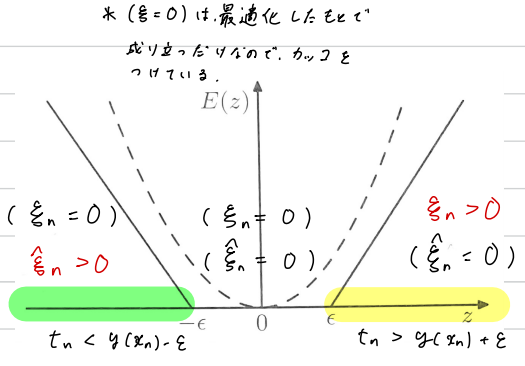
感想

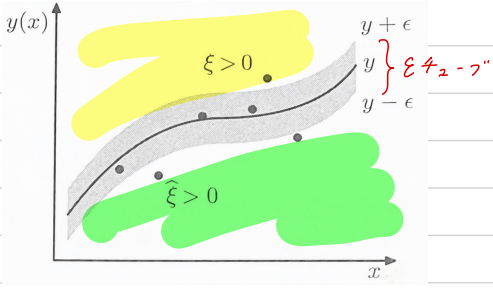
識別のSVMでは、「マージンを最大化する」という名目のもと目的関数を定めたことにより $\|w\|^2$ の項が導出されていた。一方で、回帰のSVMの誤差関数の定式化に現れる $\|w\|^2$ の説明はやや天下り的である。マージンを最大化する、というよりも、回帰モデルの正則化項由来であるという説明になっている。これはいくつかの文献を探っても同様の説明なので、悔しいがそういうものとして解釈する。

定式化を楽にするために、スラック変数 $\xi_n \geq 0$, $\hat{\xi}_n \geq 0$ を導入する。

$\xi_n > 0$ は、$t_n > y(x_n) + \varepsilon$ が成り立つデータにおいて成り立ち、

$\hat{\xi}_n > 0$ は、$t_n < y(x_n) - \varepsilon$ が成り立つデータにおいて成り立つことに注意する。





ハードマージンと同様の考えだと、**すべてのデータがεチューブの内側に入る**ことが最適化における条件となる。すなわち、 $y(x_n) - \varepsilon < t_n < y(x_n) + \varepsilon$ となる。一方で、スラック変数を導入したことにより、チューブの外側にデータが存在することを許すような条件式を構築できる。

$$\begin{cases} t_n \leq y(x_n) + \varepsilon + \xi_n \\ t_n \geq y(x_n) - \varepsilon - \hat{\xi}_n \end{cases}$$

チューブの外側のデータに存在するデータに対しては、チューブを臨時的に $\varepsilon \rightarrow \varepsilon + \xi_n$ に拡大していると解釈すると、わかりやすい。

SVMでの最適化問題は、最終的に次のように定式化される。

$$\min. \quad C\sum_{n=1}^{N}(\xi_n+\hat{\xi}_n)+\frac{1}{2}\|w^2\| \qquad (7.55)$$

$$\text{s.t.} \quad t_n \le y(x_n)+\varepsilon+\xi_n$$

$$t_n \ge y(x_n)-\varepsilon-\hat{\xi}_n$$

$$\xi_n,\ \hat{\xi}_n \ge 0$$

教科書では述べていないが、（7.52）の最小化とここでの最適化は**等価である**ことを示すことができる。最適化問題におけるいくつかの手順を準備する。

参考： http://www.msi.co.jp/nuopt/download/introduction/module/technic.pdf

準備1：絶対値最小化問題の変形

最適化問題 $\min. \sum_n |z_n|$ は、絶対値を外すために自由変数 z_n^+, z_n^- を導入すると簡単な形に直せる。z_n^+, z_n^- は次を満たす。

$$z_n = z_n^+ - z_n^-$$

$$z_n^+ \ge 0, \quad z_n^- \ge 0$$

ここで、z_n^+ は z_n が正の時の絶対値を担当し、z_n^- は z_n が負の時の絶対値を担当しているイメージである。最適化問題は次のように書ける。

$$\min. \sum_n (z_n^+ + z_n^-) \quad (= |z_n|)$$

$$\text{s.t.} \quad z_n^+ \ge 0$$

$$z_n^- \ge 0$$

準備2：最大値最小化問題の式変形

最適化問題 $\text{minimize} \ \max(z_1, z_2, \cdots, z_n)$ は次の問題と等価である。

$$\text{minimize} \quad z$$

$$\text{s.t.} \quad z \ge z_1$$

$$z \ge z_2$$

$$\vdots$$

$$z \ge z_n$$

これらを用いて、我々が初めに掲げた誤差関数の最小化問題を変形する。

$$\min. \quad C \sum_{n=1}^{N} E_\varepsilon\left(t_n - y(x_n)\right) + \frac{1}{2}\|w\|^2 \qquad (7.52)$$

$$\Leftrightarrow \quad \min. \quad C \sum_{n=1}^{N} \max\left(|t_n - y(x_n)| - \varepsilon,\ 0\right) + \frac{1}{2}\|w\|^2$$

if 文の部分を max() で表現

$$\Leftrightarrow \quad \min \quad C \sum_{n=1}^{N} \max\left(|z_n| - \varepsilon,\ 0\right) + \frac{1}{2}\|w\|^2$$

$$\Leftrightarrow \quad \min \quad C \sum_{n=1}^{N} \max\left(z_n^+ + z_n^- - \varepsilon,\ 0\right) + \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad z_n^+,\ z_n^- \geq 0$$

$$\Leftrightarrow \quad \min. \quad C \sum_{n=1}^{N} \left\{\max\left(z_n^+ - \varepsilon,\ 0\right) + \max\left(z_n^- - \varepsilon,\ 0\right)\right\} + \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad z_n^+ \cdot z_n^- \geq 0$$

$\because$ (i) $z_n^+ > 0$ のとき、$z_n^- = 0$ より、(左辺) $= \max\left(z_n^+ + 0 - \varepsilon,\ 0\right)$

(右辺) $= \max\left(z_n^+ - \varepsilon,\ 0\right) + 0$

(ii) $z_n^- > 0$ のときも同様。

(iii) $z_n^+ = z_n^- = 0$ のとき、(左辺) = (右辺) = 0

$$\Leftarrow \quad \min. \quad C \sum_{n=1}^{N} \left(\xi_n + \hat{\xi}_n\right) + \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad z_n^+,\ z_n^- \geq 0$$

$$\xi_n \geq z_n^+ - \varepsilon \quad \cdots ①$$

$$\xi_n \geq 0$$

$$\hat{\xi}_n \geq z_n^- - \varepsilon \quad \cdots ②$$

$$\hat{\xi}_n \geq 0$$

ここで: $z_n = z_n^+ - z_n^- \leq z_n^+ \quad (\because z_n^- \geq 0)$ より、

$z_n = z_n^+ - z_n^- \geq -z_n^- \quad (\because z_n^+ \geq 0)$

$$\begin{cases} z_n^+ \geq z_n \\ z_n^- \geq -z_n \end{cases}$$ が成り立つので

① $\Leftrightarrow$ $\xi_n \geq z_n^+ - \varepsilon$

$\geq z_n - \varepsilon$

$= t_n - y(x_n) - \varepsilon$

$\therefore\ t_n \leq y(x_n) + \varepsilon + \xi_n$

② $\Leftrightarrow$ $\hat{\xi}_n \geq z_n^- - \varepsilon$

$\geq -z_n - \varepsilon$

$= -t_n + y(x_n) - \varepsilon$

$\therefore\ t_n \geq y(x_n) - \varepsilon - \hat{\xi}_n$

(7.52) $\Leftrightarrow$ $\min.\ C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|w\|^2$

$S.t.\ t_n \leq y(x_n) + \varepsilon + \xi_n$

$t_n \geq y(x_n) - \varepsilon - \hat{\xi}_n$

$\xi_n \geq 0$

$\hat{\xi}_n \geq 0$

以上より、(7.52) の最小化と上記最適化問題が等価であることを示せた。

最適化問題として定式化できたので、ラグランジュの未定乗数法を適用し、双対問題の形を得る。4つの条件式それぞれに対応するラグランジュ乗数　$a_n \geq 0$, $\hat{a}_n \geq 0$, $\mu_n \geq 0$, $\hat{\mu}_n \geq 0$　を用いて、次のラグランジュ関数を構成できる。

$$L(w, b, \xi, \hat{\xi}, a, \hat{a}) = C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N}(\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) - \sum_{n=1}^{N} a_n(\varepsilon + \xi_n + y_n - t_n) - \sum_{n=1}^{N} \hat{a}_n(\varepsilon + \hat{\xi}_n - y_n + t_n) \quad (7.56)$$

$y(x_n) = w^T\phi(x_n) + b$　を代入し、 $w, b, \xi_n, \hat{\xi}_n$　での勾配が0となる条件を求めると、(7.57) - (7.60) の式が求まる。

$$w = \sum_{n=1}^{N}(a_n - \hat{a}_n)\phi(x_n) \quad (7.57)$$

$$\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0 \quad (7.58)$$

$$a_n + \mu_n = C \quad (7.59)$$

$$\hat{a}_n + \hat{\mu}_n = C \quad (7.60)$$

これらを (7.56) に代入し、双対目的関数を導く。(演習7.7)

$$L = C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)\phi^T(x_n)\phi(x_m) - \sum_{n=1}^{N}\{(C - a_n)\xi_n + (C - \hat{a}_n)\hat{\xi}_n\} - \sum_{n=1}^{N} a_n(\varepsilon + \xi_n + w\phi(x_n) + b - t_n) - \sum_{n=1}^{N}\hat{a}_n(\varepsilon + \hat{\xi}_n - w\phi(x_n) - b + t_n)$$

$-C\sum(\xi_n + \hat{\xi}_n)$

$$= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n, x_m) + \sum_{n=1}^{N}(a_n\xi_n + \hat{a}_n\hat{\xi}_n) - \sum_{n=1}^{N} a_n\xi_n - \varepsilon\sum_{n=1}^{N} a_n - \sum_{n=1}^{N} a_n\sum_{m=1}^{N}(a_m - \hat{a}_m)\phi^T(x_m)\phi(x_n) - \sum_{n=1}^{N} a_n b + \sum_{n=1}^{N} a_n t_n - \sum_{n=1}^{N}\hat{a}_n\hat{\xi}_n - \varepsilon\sum_{n=1}^{N}\hat{a}_n - \sum_{n=1}^{N}\hat{a}_n\sum_{m=1}^{N}(a_m - \hat{a}_m)\phi^T(x_m)\phi(x_n) + \sum_{n=1}^{N}\hat{a}_n b - \sum_{n=1}^{N}\hat{a}_n t_n$$

(打ち消し)

$$= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n, x_m) - \varepsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) - \sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n, x_m) - b\underbrace{\sum_{n=1}^{N}(a_n - \hat{a}_n)}_{=0\ \because (7.58)} + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n$$

$$= -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n, x_m) - \varepsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n \quad (7.61)$$

a_n, $\hat{a}_n$ に関してもソフトマージンと同じく矩形制約が存在する。

$$C \overset{(7.59)}{=} a_n + \mu_n \overset{(\mu_n \geq 0)}{\geq} a_n \qquad \geq 0$$

$\hat{a}_n$ も同様に成り立つので、まとめると、

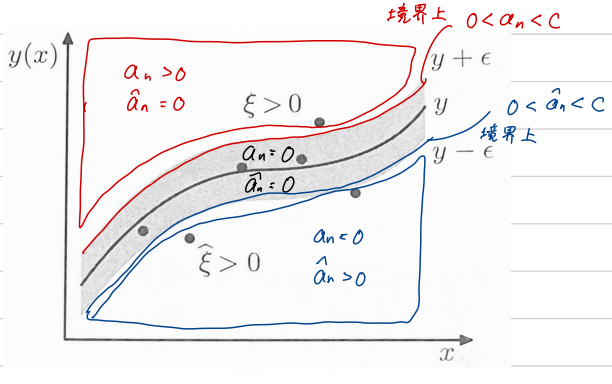
$$0 \leq a_n \leq C$$
$$0 \leq \hat{a}_n \leq C$$

これと、(7.58) が最適化における条件式となる。まとめると、次の最適化問題となる。

$$\max \quad L(a, \hat{a}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n, x_m) - \varepsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n$$

$$\text{s.t.} \quad 0 \leq a_n \leq C$$
$$0 \leq a_n \leq C$$
$$\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$$

a_n と $\hat{a}_n$ に関して解釈を与える。ε チューブの上側では $a_n > 0$ が成り立ち、下側では $\hat{a}_n > 0$ が成り立つ。ε チューブの内側では $a_n = \hat{a}_n = 0$ となる。



これらは、KKT条件の (E.11) $\lambda g(x) = 0$ から得られる (7.65) - (7.68) の条件から考察できる。導出はこれまでの条件を組み合わせるだけなので割愛する。

予測に用いるパラメータ b はチューブの境界上のデータを用いて算出できる。

$$
\begin{aligned}
b &= t_n - \varepsilon - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) \\
&= t_n - \varepsilon - \sum_{m=1}^{N}(a_m - \hat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m)
\end{aligned}
$$

最後に ν-SVM回帰について概要だけ説明する。ν-SVMでは、チューブの幅を決める ε を指定する代わりに、チューブの外側に存在するデータの割合に対する上界 ν を指定し、次の目的関数を最大化する。

$$
\begin{aligned}
\widetilde{L}(\mathbf{a}, \widehat{\mathbf{a}}) = &-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m) \\
&+ \sum_{n=1}^{N}(a_n - \widehat{a}_n)t_n.
\end{aligned} \tag{7.70}
$$

ただし，次の制約条件を仮定する．

$$0 \leqslant a_n \leqslant C/N \tag{7.71}$$

$$0 \leqslant \widehat{a}_n \leqslant C/N \tag{7.72}$$

$$\sum_{n=1}^{N}(a_n - \widehat{a}_n) = 0 \tag{7.73}$$

$$\sum_{n=1}^{N}(a_n + \widehat{a}_n) \leqslant \nu C. \tag{7.74}$$

通常のSVMでは、ε の指定のために目的変数のスケールをある程度理解して指定する必要があるが、ν-SVMではチューブ外側に含まれるデータの割合を指定するだけで良いので、パラメータの指定が幾分か楽になったように思われる。

この式が得られる過程については気になるけどパスします。