1　最大マージン分類器

1.1　ロジスティック回帰との関係

先のサブセクションにおいては、特徴量空間においてデータが線形分離不可能な場合について SVM の定式化を行なった。この定式化はロジスティック回帰モデルとの関係を議論することもできる[*1]。

ここで言う関係については次のような流れにしたがって理解されたい。

・SVM における目的関数（7.21）は既に導いた。

・この（7.21）は若干の式変形を施すことでヒンジ関数を用いた形で書き直せる（7.44-5)。

・修正した SVM の目的関数がロジスティック回帰の尤度から得られる誤差関数（7.47-8）と似た形で表現される。

ロジスティック回帰の誤差関数とヒンジ形誤差関数が似た形になることについて図 7.5 を確認し両者の関係性を視覚的に理解する。これが本レジュメでいうところの「関係」の意味である。本サブセクションにおいては、この関係について数式を追いかけながら理解してもらうことにする。

その後にロジスティック回帰の誤差関数やヒンジ形誤差関数を用いる利点を抑える。

【ロジスティック回帰の誤差関数とヒンジ形誤差関数】

まずは先のサブセクションで導いた（7.21）がヒンジ形誤差関数で書き表せることを示す。

$$
\begin{aligned}
(7.21) &= C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|\boldsymbol{w}\|^2 \\
&= C\sum_{n=1}^{N}[1-y_nt_n]_+ + \frac{1}{2}\|\boldsymbol{w}\|^2 \ (\because y_nt_n \geq 1 \text{ ならば} \xi_n = 0 \text{ ; otherwise } \xi_n = 1-y_nt_n) \\
&\sim \sum_{n=1}^{N} E_{SV}(y_nt_n) + \lambda\|\boldsymbol{w}\|^2 \ (\because E_{SV}(y_nt_n) = [1-y_nt_n]_+,\ \lambda = (2C)^{-1}) \qquad (1)
\end{aligned}
$$

ここで $E_{SV}(y_nt_n) = [1-y_nt_n]_+$ はヒンジ関数と呼ばれる。この関数はデータが正しく分類される場合は損失がゼロ、誤分類される場合はペナルティとして $\xi_n(>0)$ を与えるように定義されている。ここで導いた関数のグラフ形についてはこの後で触れることにするので、ひとまずスキップさせてもらう。

次にロジスティック回帰についての目的関数を尤度関数と正則化項の定義から導出する。今知りたい尤度関数は式（4.87)[*2]を用いて表すことができる。ただしクラスラベルを $\{0,1\} \to \{-1,1\}$ と定義し直すことに注意せよ。また Given な量は $y(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{\phi}(\boldsymbol{x}) + b$ で約束される。

これらの前提のもとでクラス 1 と -1 に分類される確率を $p(t=1 \mid y), p(t=-1 \mid y)$ と定義するとそれぞれロジスティックシグモイド関数を用いて次のように書き表すことができる。

$$
p(t=1 \mid y) = \sigma(y),\ p(t=-1 \mid y) = 1-\sigma(y) = \sigma(-y) \qquad (2)
$$

これは t と y の符号の関係から結局

$$
p(t \mid y) = \sigma(yt) \qquad (3)
$$

[*1] 上巻頁 204 において特徴量が与えられた時のクラスの事後確率が定式化されている。この事後確率はロジスティックシグモイド関数で与えられることが知られている。このモデルのことをロジスティック回帰モデルと呼ぶのであった。

[*2] 上巻 4 章頁 204 中段を見よ。

と書き表すことができる。添字 n を露わに書くと $p(t_n \mid y_n) = \sigma(y_n t_n)$ となる。各 n が独立であると仮定すると尤度関数は式 3 の積で書くことができる。これに対数を噛ませることで対数尤度は次のように表せる。

$$(\text{尤度}) = \sum_n^N \ln p(t_n \mid y_n) = \sum_n^N \ln \sigma(y_n t_n) \tag{4}$$

これに 2 次の正則化を加えることによって目的関数は

$$\sum_n^N E_{LR}(y_n t_n) + \lambda \|\boldsymbol{w}\|^2 \tag{5}$$

と求められる。ここで $E_{LR}(y_n t_n) = \ln \sigma(y_n t_n) = \ln\left(1 + \exp(-y_n t_n)\right)$ と約束している。

この $E_{LR}(y_n t_n)$ と先に求めたヒンジ関数 $E_{SV}(y_n t_n)$ と比較するために前者を $\ln(2)$ で割り算しておく。こうすることで $E_{LR}(y_n t_n)$ は $y_n t_n$ を 0 としたときに平面上の座標 $(0, 1)$ を通るようになる。このようにスケールを調整した関数のグラフを図示すると図 7.5 のようになる。図からわかるようにスケール調整した $E_{LR}(y_n t_n)$ がヒンジ関数 $E_{SV}(y_n t_n)$ と似た形になっていることがわかる。ただし違いとして SVM の方では $y_n t_n \geq 1$ となる領域において疎な解が得られることがある。

そのほかの誤差関数の選び方としては例えば二乗和誤差関数がある。二乗和誤差関数を選んだ時のデメリットについては読み合わせ。

1.2 多クラス SVM

ここまで見てきた SVM の手法は 2 クラス分類問題を対象としたものであった。このサブセクションでは多クラス分類問題を取り扱うための手法について近年提案されているものを紹介していく。

ここでは次の 3 つの手法について基本的な考え方や pros/cons をマトリックスで整理する。

1. one-versus-the-rest
2. one-versus-one
3. 誤り訂正出力符号

そして最後にラベル付による分類問題からは外れて、確率密度分布の推定問題と関連する教師なし学習問題を解く手法（単一クラス SVM）を紹介する。

表 1　-多クラス SVM-各手法の pros/cons 比較、総クラス数は K 個

手法	one-versus-the-rest	one-versus-one	誤り訂正出力符号
考え方	C_k に属するデータを正例 それ以外のデータを負例	全てのクラスを 組み合わせて学習	←と同じ?
分類器の個数	$K - 1$	$\frac{K(K-1)}{2}$	???
問題点	曖昧な分類領域 解の妥当性× 解の対称性×	曖昧な分類領域 学習時間長	???
計算負荷 (学習時)	$O(KN^2)$	$O(K^2N^2)$	???

【表について補足】

・曖昧な分類領域のイメージについては上巻頁 181 図 4.2 を確認

・one-versus-the-rest のクラス割り当てについて一つの入力に複数のクラスがラベル付けされることも生じうる。その際 $y(\boldsymbol{x}) = \max_k y_k(\boldsymbol{x})$ となるようなクラスを予測値とすることがある。しかし、個々の SVM が独立な問題を解いていることから $y_k(\boldsymbol{x})$ の値を比較することに意味があるかどうかは保証されていない。

・one-versus-the-rest において対称性が失われるのは個々の分類器では正例と負例のデータ数のバランスが悪くなってしまうことによる。例えば 10 クラス分類問題を考える。このときどのクラスも同じデータ数を持っている場合、正例と負例の割合が $1:9$ になってしまう。このような偏った訓練データで学習をすることは望ましくない。

・誤り訂正出力符号は意味不明でした。すいません。

ここまで各手法の考え方や問題点などを整理してきた。しかしながら、オチとしては多クラス分類問題への SVM 適用は今なお未解決問題ということである。ちなみに、今のところ広く用いられている手法は one-versus-the-rest である[*3]。

単一クラス SVM については読み合わせ。

今回は誤差関数がロジスティック回帰の誤差関数やヒンジ形誤差関数で与えられる場合における SVM 手法を取り扱った。次回は誤差関数が二乗和誤差の形で与えられる場合において SVM を適用する手法を学ぶ。ただし疎な解を持つように若干の修正を加えることになることに注意されたい。

[*3] なぜこの手法を使うのか理由は明記されていないが、やはり one-versus-the-rest の方が計算負荷が one-versus-one に比べて低いからだと勝手に理解している。