決定的二分機械の学習と予測誤差

渡辺 一帆
東京工業大学大学院 総合理工学研究科 電子機能システム専攻

平成 14 年 12 月 6 日

1 決定的二分機械

単純パーセプトロンやサポートベクターマシンなどの二分機械について考える。ノイズのない決定的二分機械は、入力 $\mathbf{x} \in R^m$ に対して、パラメータ $\mathbf{w}$ をもつ関数 f を用いて、出力 $y \in \{+1, -1\}$ を、

$$y = \mathrm{sign} f(\mathbf{x}, \mathbf{w})$$

により計算する。

以下では、簡単のために $\mathbf{w} \in R^m$ として、

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x}$$

である単純パーセプトロンの場合を考える。

このとき、パラメータ $\mathbf{w}$ をもつ機械の、入力 $\mathbf{x}$ が与えられたときの出力 y の分布は、

$$p(y|\mathbf{x}, \mathbf{w}) = \begin{cases} 1 & y(\mathbf{w}^T \mathbf{x}) > 0 \\ 0 & \text{それ以外} \end{cases}$$

で与えられる。

真の機械が存在するとし、パラメータ $\mathbf{w}_0$ をもつとする。入力 $\mathbf{x}$ の分布 $p(\mathbf{x})$ と真の分布 $p(y|\mathbf{x}, \mathbf{w}_0)$ から発生された t 個の学習データ

$$D^{(t)} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_t, y_t)\}$$

を用いて学習をおこなう。図 1 に示すように、t 個の学習データに正解するパラメータは無数に存在する。このことをパラメータ空間で見ると図 2 のようになり、このような学習データ全てに正解するパラメータの領域を許容領域といい、A_t で表す。すなわち、

$$A_t = \{\mathbf{w} | y_i(\mathbf{w}^T \mathbf{x}) > 0, i = 1, \cdots, t\}$$

2 学習アルゴリズム

学習アルゴリズムとして以下のものを考える。

(i). **Gibbs アルゴリズム**

パラメータの事前分布 $p(\mathbf{w})$ が与えられたとする。

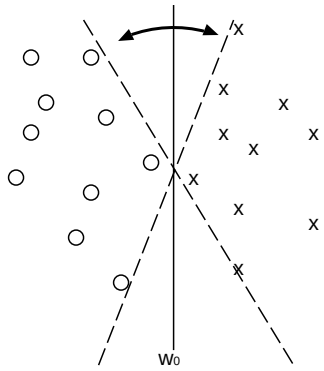

図 1: 入力空間におけるパラメータの広がり

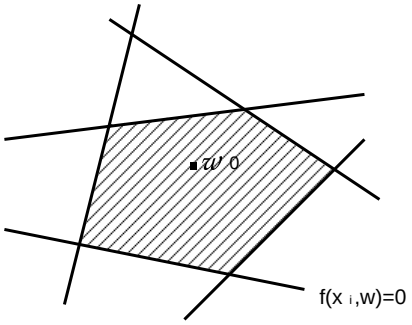

図 2: 許容領域

パラメータの事後分布を

$$\begin{aligned} p(\mathbf{w}|D^{(t)}) &= \frac{p(\mathbf{w})}{Z_t}\prod_{i=1}^{t} p(y_i|\mathbf{x},\mathbf{w}) \\ &= \begin{cases} \frac{p(\mathbf{w})}{Z_t} & \mathbf{w}\in A_t \\ 0 & \text{その他} \end{cases} \end{aligned}$$

により構成する。ここで、

$$Z_t = \int_{A_t} p(\mathbf{w})d\mathbf{w}$$

である。事後分布に従ってパラメータ $\hat{\mathbf{w}}_t$ を一つ選び、推定パラメータとする。

(ii). **Bayes アルゴリズム**

パラメータの推定は行わずに、上の事後分布を用いて、テスト入力に対する予測を直接行う。

$$\begin{cases} \hat{y}=1 & \int_{A+} p(\mathbf{w}|\mathbf{D^{(t)}})d\mathbf{w} > \frac{1}{2} \\ \hat{y}=-1 & \text{その他} \end{cases}$$

ここで、

$$A_+ = \{\mathbf{w}|f(\mathbf{x}_{t+1},\mathbf{w})>0, \mathbf{w}\in A_t\} \quad (\mathbf{x}_{t+1}\text{ はテスト入力})$$

(iii). **最悪アルゴリズム**　(予測誤差を考察するためのみ)

パラメータ空間で、$f(\mathbf{x}_{t+1},\mathbf{w})=0$ (テスト入力が作る超平面) が、許容領域 A_t と交わるときには、$\hat{\mathbf{w}}_t$ の選び方によっては予測に失敗する。このとき、常にわざと失敗するように選ぶ。

3　平均予測誤差

予測誤差 $e(D^t)$ は 1 つのテスト入力に対するパーセプトロンの出力が間違う確率として定義される。
平均予測誤差 $e(t)$ は、学習データの出方について平均をとることで、

$$e(t) = E_{D^{(t)}}[e(D^{(t)})]$$

である。

(i). **Gibbs アルゴリズム**

$$\text{Prob}\{(\mathbf{x}_{t+1}, y_{t+1}) \text{ に間違う }\}$$

$$
\begin{aligned}
&= 1 - \mathrm{Prob}\{(\mathbf{x}_{t+1}, y_{t+1}) \text{ に正解 }\} \\
&= 1 - \int p(y_{t+1}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|D^{(t)}) d\mathbf{w} \\
&= 1 - \frac{\int_{A_t} p(y_{t+1}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}}{Z_t} \\
&= 1 - \frac{\int_{A_{t+1}} p(\mathbf{w}) d\mathbf{w}}{Z_t} \\
&= 1 - \frac{Z_{t+1}}{Z_t}
\end{aligned}
$$

より、

$$
\begin{aligned}
\text{予測誤差} \quad e_G(D^{(t)}) &= E_{\binom{\mathbf{x}_{t+1}}{y_{t+1}}}\left[1 - \frac{Z_{t+1}}{Z_t}\right] \\
\text{平均予測誤差} \quad e_G(t) &= E_{D^{(t+1)}}\left[1 - \frac{Z_{t+1}}{Z_t}\right]
\end{aligned}
$$

(ii). **Bayes アルゴリズム**

$$
\begin{aligned}
&\text{"}\hat{y}(\text{予測結果}) \neq y_{t+1}\text{"} \\
\Leftrightarrow \quad &\begin{cases} \hat{y} = -1 \ \text{ かつ } \ y_{t+1} = 1 \\ \hat{y} = 1 \ \text{ かつ } \ y_{t+1} = -1 \end{cases} \\
\Leftrightarrow \quad &\begin{cases} \int_{A_+} p(\mathbf{w}|D^{(t)}) d\mathbf{w} < \frac{1}{2} \ \text{ かつ } \ y_{t+1} = 1 \\ 1 - \int_{A_+} p(\mathbf{w}|D^{(t)}) d\mathbf{w} < \frac{1}{2} \ \text{ かつ } \ y_{t+1} = -1 \end{cases} \\
\Leftrightarrow \quad &\int_{A_{t+1}} p(\mathbf{w}|D^{(t)}) d\mathbf{w} < \frac{1}{2} \\
\Leftrightarrow \quad &\int_{A_{t+1}} \frac{p(\mathbf{w})}{Z_t} d\mathbf{w} < \frac{1}{2} \\
\Leftrightarrow \quad &\frac{Z_{t+1}}{Z_t} < \frac{1}{2}
\end{aligned}
$$

より、平均予測誤差 e_B は、

$$
e_B(t) = E_{D^{(t+1)}}\left[\Theta\left(\frac{Z_{t+1}}{Z_t} < \frac{1}{2}\right)\right] \quad (\Theta(\cdot) \text{ はインジケータ関数})
$$

(iii). **最悪アルゴリズム**

学習データ $D^{(t)}$ から、ある例題 $(\mathbf{x}_i, y_i)$ を除いたときに、許容領域 A_t が変化するとき、その例題を有効例題と呼ぶ。A_t の有効例題数の期待値を F_t とする。

最悪アルゴリズムでは、$f(\mathbf{x}_{t+1}, \mathbf{w}) = 0$ が A_t と交わるとき、必ず予測を誤る。このとき、A_{t+1} では $(\mathbf{x}_{t+1}, y_{t+1})$ が有効例題となることから、最悪アルゴリズムの平均予測誤差 e_w は、

$$
e_w(t) = \frac{F_{t+1}}{t+1}
$$

となる。

4 予測エントロピー

Gibbs アルゴリズムの予測誤差は、

$$
e_G(D^{(t)}) = E_{\binom{\mathbf{x}_{t+1}}{y_{t+1}}}\left[1 - \frac{Z_{t+1}}{Z_t}\right]
$$

であった。しかし、これは漸近的な場合でも計算が困難なので、次式によって対数予測誤差を定義する。

$$e^*(D^{(t)}) = E_{\binom{x_{t+1}}{y_{t+1}}}\Big[-\log\frac{Z_{t+1}}{Z_t}\Big]$$

$1-x \leq -\log x$ なので、

$$e_G(D^{(t)}) \leq e^*(D^{(t)})$$

よって、対数予測誤差は予測誤差の上限を与える。

対数予測誤差の平均により、予測エントロピー $e^*(t)$ を定義する。

$$\begin{aligned} e^*(t) &= E_{D^{(t)}}\Big[e^*(D^{(t)})\Big] \\ &= E_{D^{(t+1)}}\Big[-\log\frac{Z_{t+1}}{Z_t}\Big] \\ &= E_{D^{(t)}}\Big[\log Z_t\Big] - E_{D^{(t+1)}}\Big[\log Z_{t+1}\Big] \end{aligned}$$

すると、

$$e_G(t) \leq e^*(t)$$

が成り立つ。

以下では予測エントロピーの性質について調べる。

補題

$$Y_t = t^m Z_t \qquad t = 1, 2, \cdots$$

とおくと、Y_t はある確率変数 Y に法則収束する。

(証明略)

上の補題を用いると、

$$\begin{aligned} E_{D^{(t)}}\Big[\log Z_t\Big] &\simeq E_{D^{(t)}}\Big[\log(t^{-m}Y)\Big] \\ &= E_{D^{(t)}}\Big[\log Y\Big] - m\log t \end{aligned}$$

よって、

$$\begin{aligned} e^*(t) &= E_{D^{(t)}}\Big[\log Z_t\Big] - E_{D^{(t+1)}}\Big[\log Z_{t+1}\Big] \\ &\simeq m\log(t+1) - m\log t \\ &= \frac{m}{t} + O\big(\frac{1}{t^2}\big) \end{aligned}$$

より、予測エントロピーの漸近特性は、

$$e^*(t) = \frac{m}{t}$$

となることがわかる。

参考文献

[1] K.Ikeda and S.Amari "Geometry of admissible parameter region in neural learning", IEICE Trans.Fundamentals,E79-A,1996,pp.938-943.

[2] 池田和司,"多項式カーネルを持つカーネル法の学習曲線", Proc.IBIS2002,pp.25-30,2002

[3] 村田昇,"学習の統計的漸近理論"