

機械学習のための数学

Mathematics for Machine Learning

杉山 将 (東京大学)

Masashi Sugiyama (University of Tokyo)

sugi@k.u-tokyo.ac.jp, <http://www.ms.k.u-tokyo.ac.jp>

鈴木 大慈 (東京工業大学)

Taiji Suzuki (Tokyo Institute of Technology)

s-taiji@is.titech.ac.jp

概要

インターネットから得られるテキスト, 音声, 画像, 映像などのマルチメディアデータ, ウェアラブル機器から得られるライフログデータ, パーソナルファブ리케이션データ, ロボットの探査データなど, 様々な場面で膨大な量のデータが収集・蓄積されるようになってきた. このようなビッグデータを活用して新たな価値を創造するためには, 機械学習による知的なデータ解析が重要である. 本稿では, 統計的パターン認識を例に取り, 機械学習でよく用いられる数学の基礎的事項を解説する.

1 統計的パターン認識の枠組み

パターン認識の目的は, パターン $\mathbf{x} \in \mathbb{R}^d$ をそれが属するクラス $y \in \{1, \dots, c\}$ に割り当てることである. ここで, \mathbb{R}^d は d 次元の実ベクトルを表し, c はクラス数を表す. 例えば, 16×16 画素の画像に書かれた手書き数字を認識する問題では, 次元数は $d = 16 \times 16 = 256$ であり, クラス数は数字の $0 \sim 9$ に対応して $c = 10$ である.

統計的パターン認識では, パターン \mathbf{x} が従う確率 $p(\mathbf{x})$, クラス y が従う確率 $p(y)$, それらの同時確率 $p(\mathbf{x}, y)$, パターン \mathbf{x} がクラス y に属する条件付き確率 $p(y|\mathbf{x})$, クラス y に属するパターン \mathbf{x} の条件付き確率 $p(\mathbf{x}|y)$ を考える¹. $p(y|\mathbf{x})$ はパターン \mathbf{x} を見た後でのクラス y の出現確率, $p(y)$ はパターン \mathbf{x} を見る前のクラス y の出現確率を表すことから, これらをそれぞれクラス y の事後確率, 事前確率とよぶ. 事後確率 $p(y|\mathbf{x})$ が最大と

¹正確には, $p(\mathbf{x})$ は確率密度関数, $p(y)$ は確率質量関数であるが, 表記を簡単にするため, 本稿ではこれらを単に「確率」とよぶことにする.

なるクラス y にパターン \mathbf{x} を分類すれば、認識誤差が最小になるため、これが理論的に最適なパターン認識法である。

しかし実際には事後確率 $p(y|\mathbf{x})$ が未知であるため、同時確率 $p(\mathbf{x}, y)$ に独立に従う n 個の訓練標本 $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ を用いてパターン認識器を構成することにする。 $p(y|\mathbf{x})$ を訓練標本から直接推定する方式を識別的アプローチ [5, 4] とよぶ。「識別的」という名称は、パターン \mathbf{x} をクラス y に識別するために必要な事後確率 $p(y|\mathbf{x})$ そのものを直接推定することに由来している。一方、ベイズの定理

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \propto p(\mathbf{x}|y)p(y) \quad (1)$$

を用いれば、 $p(y|\mathbf{x})$ の y に関する最大化を $p(\mathbf{x}|y)p(y)$ の最大化に変換できる。ここで、 \propto は比例関係を表す。 $p(\mathbf{x}|y)p(y)$ を訓練標本から推定する事によってパターン認識を行う方式を、生成的アプローチ [1, 3] とよぶ。「生成的」という名称は、データの生成確率 $p(\mathbf{x}|y)p(y) = p(\mathbf{x}, y)$ を推定することに由来している。

2 生成的アプローチ

生成的アプローチでは、パラメータ θ_y を持つパラメトリックモデル $q(\mathbf{x}|\theta_y)$ を考え、クラス y に属する訓練パターン数 n_y を用いて単純に $\hat{p}(y) = n_y/n$ と近似することが多い。一方、 $p(\mathbf{x}|y)$ の推定には様々な方法が用いられる。

2.1 最尤推定法

最尤推定法では、パラメータ θ_y を持つパラメトリックモデル $q(\mathbf{x}|\theta_y)$ を考え、クラス y に属する訓練パターン $\mathcal{X}_y = \{\mathbf{x}_i\}_{i:y_i=y}$ が生成される確率の対数である対数尤度

$$\log p(\mathcal{X}_y|\theta_y) = \sum_{i:y_i=y} \log q(\mathbf{x}_i|\theta_y)$$

を最大にするようにパラメータ θ_y を決定する。すなわち、 $p(\mathbf{x}|y)$ の推定量 $\hat{p}(\mathbf{x}|y)$ は

$$\hat{p}(\mathbf{x}|y) = q(\mathbf{x}|\hat{\theta}_y), \quad \hat{\theta}_y = \operatorname{argmax}_{\theta_y} \log p(\mathcal{X}_y|\theta_y)$$

で与えられる。「最尤推定法」という名称は、与えられた訓練パターン $\mathcal{X}_y = \{\mathbf{x}_i\}_{i:y_i=y}$ のもとで、最も尤もらしいパラメータ値を求めることに由来している。

例えば、パラメトリックモデル $q(\mathbf{x}|\theta_y)$ として、期待値ベクトル $\boldsymbol{\mu}_y$ と分散共分散行列 $\boldsymbol{\Sigma}_y$ をパラメータとするガウスモデル

$$q_G(\mathbf{x}|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_y|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y)\right) \quad (2)$$

を用いることにしよう。ここで、 $|\cdot|$ は行列式を表す。対数尤度 $\sum_{i:y_i=y} \log q_G(\mathbf{x}_i|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ の微分をゼロとおけば、最尤推定量 $\hat{\boldsymbol{\mu}}_y, \hat{\boldsymbol{\Sigma}}_y$ は、

$$\hat{\boldsymbol{\mu}}_y = \frac{1}{n_y} \sum_{i:y_i=y} \mathbf{x}_i, \quad \hat{\boldsymbol{\Sigma}}_y = \frac{1}{n_y} \sum_{i:y_i=y} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_y)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_y)^\top$$

と解析的に求められる。

全クラス $y = 1, \dots, c$ の分散共分散行列 $\boldsymbol{\Sigma}_y$ が等しいと仮定すると、その共通の分散共分散行列 $\boldsymbol{\Sigma}$ の最尤推定量 $\hat{\boldsymbol{\Sigma}}$ は

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{y=1}^c \sum_{i:y_i=y} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_y)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_y)^\top = \sum_{y=1}^c \frac{n_y}{n} \hat{\boldsymbol{\Sigma}}_y$$

で与えられる。このとき、クラス y の対数事後確率 $\log p(y|\mathbf{x})$ は

$$\log \hat{p}(y|\mathbf{x}) = \mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_y - \frac{1}{2} \hat{\boldsymbol{\mu}}_y^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_y + \log \frac{n_y}{n} + C$$

と近似できる。ここで、 C は y に依存しない定数である。従って、クラス間の分離境界は \mathbf{x} の一次形式となる。これをフィッシャーの判別分析とよぶ。

2.2 ベイズ推定法

最尤推定法ではモデルのパラメータの値 $\boldsymbol{\theta}_y$ を訓練標本 \mathcal{X}_y から推定した。一方、ベイズ推定法ではパラメータ $\boldsymbol{\theta}_y$ を確率変数とみなして、パラメータの値そのものではなくその事後確率 $p(\boldsymbol{\theta}_y|\mathcal{X}_y)$ を求める。具体的には、パラメータ $\boldsymbol{\theta}_y$ の事前確率 $p(\boldsymbol{\theta}_y)$ を考え、ベイズの定理

$$p(\boldsymbol{\theta}_y|\mathcal{X}_y) \propto p(\mathcal{X}_y|\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_y)$$

を用いて事後確率 $p(\boldsymbol{\theta}_y|\mathcal{X}_y)$ を計算する。「ベイズ推定法」という名称は、ベイズの定理を用いて事後確率を計算することに由来している。

事後確率 $p(\boldsymbol{\theta}_y|\mathcal{X}_y)$ の形状はモデル $q(\mathbf{x}|\boldsymbol{\theta}_y)$ と事前確率 $p(\boldsymbol{\theta}_y)$ に依存し、組み合わせによっては事後確率の計算が困難になることがある。もし、モデル $q(\mathbf{x}|\boldsymbol{\theta}_y)$ と共役な事前確率 $p(\boldsymbol{\theta}_y)$ を用いれば、事後確率 $p(\boldsymbol{\theta}_y|\mathcal{X}_y)$ は事前確率と同じ種類の確率分布になり、計算が容易になる。例えば、式(2)で示した期待値ベクトル $\boldsymbol{\mu}_y$ と分散共分散行列 $\boldsymbol{\Sigma}_y$ をパラメータとするガウスモデルに対しては、正規・逆ウィシャート分布

$$p(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \propto |\boldsymbol{\Sigma}_y|^{-\frac{\nu+d+2}{2}} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{S}\boldsymbol{\Sigma}_y^{-1})\right) \exp\left(-\frac{\kappa}{2} (\boldsymbol{\mu}_y - \mathbf{m})^\top \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{\mu}_y - \mathbf{m})\right)$$

が共役な事前確率である。ここで、 \mathbf{m} は d 次元ベクトル、 κ は正のスカラー、 ν は正の整数、 \mathbf{S} は正定値対称行列であり、これらは正規・逆ウィシャート分布のパラメータである。事後確率では、これらのパラメータは

$$\begin{aligned}\mathbf{m} &\leftarrow \frac{\kappa \mathbf{m} + n_y \hat{\boldsymbol{\mu}}_y}{\kappa + n_y}, \quad \kappa \leftarrow \kappa + n_y, \quad \nu \leftarrow \nu + n_y, \\ \mathbf{S} &\leftarrow \mathbf{S} + n_y \hat{\boldsymbol{\Sigma}}_y + \frac{\kappa n_y}{\kappa + n_y} (\hat{\boldsymbol{\mu}}_y - \mathbf{m})(\hat{\boldsymbol{\mu}}_y - \mathbf{m})^\top\end{aligned}$$

となる。

パラメータの事後確率 $p(\boldsymbol{\theta}_y | \mathcal{X}_y)$ を求めた後は、モデル $q(\mathbf{x} | \boldsymbol{\theta}_y)$ の事後確率に関する期待値

$$\bar{p}(\mathbf{x} | y) = \int q(\mathbf{x} | \boldsymbol{\theta}_y) p(\boldsymbol{\theta}_y | \mathcal{X}_y) d\boldsymbol{\theta}_y \quad (3)$$

を $p(\mathbf{x} | y)$ の推定結果とする。これをベイズ予測分布とよぶ。最尤推定では、モデル $q(\mathbf{x} | \boldsymbol{\theta}_y)$ に最尤推定量 $\hat{\boldsymbol{\theta}}_y$ を代入することにより確率分布の推定量 $q(\mathbf{x} | \hat{\boldsymbol{\theta}}_y)$ を求めた。従って、例えば式 (2) で示したガウスモデルを用いれば、推定量 $q(\mathbf{x} | \hat{\boldsymbol{\theta}}_y)$ もガウス分布である。一方、ベイズ予測分布はモデル $q(\mathbf{x} | \boldsymbol{\theta}_y)$ を平均化するため、例えばガウスモデルを用いたとしても式 (3) は一般にガウス分布にはならない。このように、パラメトリックモデルを用いるにもかかわらず、モデルで表すことのできない確率分布も表現できるのが、ベイズ予測分布の特徴である。

しかし、式 (3) の積分も一般に計算が困難なため、変分法やサンプリングとよばれる技法による近似がよく用いられる。また最大事後確率推定法では、対数事後確率 $\log p(\boldsymbol{\theta}_y | \mathcal{X}_y)$ を最大にするパラメータ値 $\tilde{\boldsymbol{\theta}}_y$ を用いて、

$$\tilde{p}(\mathbf{x} | y) = q(\mathbf{x} | \tilde{\boldsymbol{\theta}}_y), \quad \tilde{\boldsymbol{\theta}}_y = \underset{\boldsymbol{\theta}_y}{\operatorname{argmax}} \log p(\boldsymbol{\theta}_y | \mathcal{X}_y)$$

と $p(\mathbf{x} | y)$ を推定する。 $\tilde{\boldsymbol{\theta}}_y$ は、

$$\tilde{\boldsymbol{\theta}}_y = \underset{\boldsymbol{\theta}_y}{\operatorname{argmax}} \log p(\mathcal{X}_y | \boldsymbol{\theta}_y) + \log p(\boldsymbol{\theta}_y)$$

とも書けることから、最大事後確率推定法では事前確率による罰則項 $\log p(\boldsymbol{\theta}_y)$ を対数尤度に加えた量を最大化していることがわかる。そのため、最大事後確率推定法は罰則付き最尤推定法とよばれることもある。実際、ベイズ予測分布と異なり、最大事後確率推定法はモデルの中でのみ確率分布を推定するため、ベイズ推定よりは最尤推定に性質が近いと言える。

2.3 カーネル密度推定法

最尤推定法やベイズ推定法では、 $p(\mathbf{x} | y)$ の推定にパラメトリックモデル $q(\mathbf{x} | \boldsymbol{\theta}_y)$ を用いた。一方ノンパラメトリック法では、パラメトリックモデルを用いずに $p(\mathbf{x} | y)$ を直接推定する。

代表的なノンパラメトリック法であるカーネル密度推定法では、

$$\hat{p}(\mathbf{x}|y) = \frac{1}{n_y} \sum_{i:y_i=y} K(\mathbf{x}, \mathbf{x}_i)$$

によって $p(\mathbf{x}|y)$ を推定する。ここで、 $K(\mathbf{x}, \mathbf{x}')$ はカーネルとよばれる 2 変数関数であり、ガウスクーネル

$$K(\mathbf{x}, \mathbf{x}') = (2\pi h^2)^{-\frac{d}{2}} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2h^2}\right)$$

がよく用いられる。 $h > 0$ はバンド幅とよばれ、ガウスクーネルの滑らかさを調整するパラメータである。バンド幅は、例えば以下の交差確認法によって決定する。

1. クラス y に属する訓練パターン $\mathcal{X}_y = \{\mathbf{x}_i\}_{i:y_i=y}$ を $\mathcal{X}_y^{(1)}, \dots, \mathcal{X}_y^{(T)}$ に等分割する。
2. $\mathcal{X}_y^{(t)}$ 以外の訓練パターンを用いて、バンド幅 h のカーネル密度推定量 $\hat{p}_h^{(t)}(\mathbf{x}|y)$ を求め、 $\mathcal{X}_y^{(t)}$ での平均対数尤度を求める。

$$L_h^{(t)} = \frac{1}{|\mathcal{X}_y^{(t)}|} \sum_{\mathbf{x}^{(t)} \in \mathcal{X}_y^{(t)}} \log \hat{p}_h^{(t)}(\mathbf{x}^{(t)}|y)$$

3. これを $t = 1, \dots, T$ に対して実行し、平均対数尤度の平均 $L_h = \frac{1}{T} \sum_{t=1}^T L_h^{(t)}$ を求める。
4. これを様々なバンド幅 h に対して実行し、 L_h を最大にするものを選択する。

2.4 最近傍密度推定法

最近傍密度推定法は、クラス y に属する訓練パターン $\mathcal{X}_y = \{\mathbf{x}_i\}_{i:y_i=y}$ の中で注目点 \mathbf{x} から k 番目に近い点 $\mathbf{x}^{(k)}$ への距離 $r_y(k) = \|\mathbf{x}^{(k)} - \mathbf{x}\|$ を用いて、

$$\hat{p}(\mathbf{x}|y) = \frac{k\Gamma(\frac{d}{2} + 1)}{n_y \pi^{\frac{d}{2}} r_y(k)^d}$$

によって $p(\mathbf{x}|y)$ を推定するノンパラメトリック法である。ここで、

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$$

はガンマ関数であり、階乗 $\alpha!$ の実数への一般化になっている。

$k = 1$ の最近傍密度推定法を式 (1) に適用すれば、事後確率 $p(y|\mathbf{x})$ の近似の y に関する最大化は、 $r_y(1)$ の y に関する最小化に帰着される：

$$\operatorname{argmax}_y \hat{p}(\mathbf{x}|y)\hat{p}(y) = \operatorname{argmax}_y \frac{\Gamma(\frac{d}{2} + 1) n_y}{n_y \pi^{\frac{d}{2}} r_y(1)^d n} = \operatorname{argmax}_y r_y(1)^{-d} = \operatorname{argmin}_y r_y(1)$$

すなわち、全クラスの訓練パターン $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ の中で注目点 \mathbf{x} に最も近いパターンが属するクラス y を、 \mathbf{x} が属するクラスと予測する。これを最近傍識別器とよぶ。これを一般化して、注目点 \mathbf{x} の近傍 k 個のパターンが属するクラスの多数決を取る方式を、 k 最近傍識別器とよぶ。近傍数 k の値は、例えば上述した交差確認法によって決定できる。

3 識別的アプローチ

識別的アプローチでは、クラス y の事後確率 $p(y|\mathbf{x})$ を直接推定する。

3.1 ロジスティック回帰

ロジスティック回帰では、事後確率 $p(y|\mathbf{x})$ をカーネルロジスティックモデル

$$q_L(y|\mathbf{x}, \boldsymbol{\theta}) = \frac{\exp\left(\sum_{j=1}^n \theta_j^{(y)} K(\mathbf{x}, \mathbf{x}_j)\right)}{\sum_{y'=1}^c \exp\left(\sum_{j'=1}^n \theta_{j'}^{(y')} K(\mathbf{x}, \mathbf{x}_{j'})\right)}$$

によってモデル化する。そして、最尤推定法やベイズ推定法によって、パラメータ $\{\theta_j^{(y)}\}_{y=1, j=1}^{c, n}$ を訓練標本から学習する。

ロジスティック回帰の最尤推定量は、例えば勾配法によって求められる。すなわち、パラメータ $\boldsymbol{\theta}$ を適当な初期値に設定し、対数尤度の勾配を上昇するようにパラメータ値を更新する：

$$\theta_j^{(y)} \leftarrow \theta_j^{(y)} + \varepsilon \sum_{i=1}^n \left(I(y = y_i) K(\mathbf{x}_i, \mathbf{x}_j) - \frac{\exp\left(\sum_{j'=1}^n \theta_{j'}^{(y)} K(\mathbf{x}_i, \mathbf{x}_{j'})\right) K(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{y''=1}^c \exp\left(\sum_{j''=1}^n \theta_{j''}^{(y'')} K(\mathbf{x}_i, \mathbf{x}_{j''})\right)} \right)$$

for $y = 1, \dots, c, j = 1, \dots, n$

ここで、 $I(y = y_i)$ は $y = y_i$ のとき 1、 $y \neq y_i$ のとき 0 を出力する指示関数であり、 $\varepsilon > 0$ は勾配上昇のステップ幅を表す。この勾配上昇をパラメータ値が収束するまで繰り返すことにより、最尤推定量が求められる。なお、上記の勾配の計算において、 n 個の訓練標本全てを用いずにランダムに選んだ部分集合（ミニバッチ）で近似することにより、収束速度を改善できることがある。

ここで、2クラスの分類問題を考え $y = \pm 1$ とおき、パラメータ $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^\top$ を持つカーネルモデル

$$f(\mathbf{x}) = \sum_{j=1}^n \theta_j K(\mathbf{x}, \mathbf{x}_j) \quad (4)$$

を用いて、ロジスティックモデルを

$$q_L(y = +1|\mathbf{x}) = \frac{\exp(f(\mathbf{x}))}{\exp(f(\mathbf{x})) + \exp(-f(\mathbf{x}))}$$

$$q_L(y = -1|\mathbf{x}) = 1 - q_L(y = +1|\mathbf{x})$$

と表現することにする。そうすると、ロジスティックモデルの最尤推定は

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^n \log(1 + \exp(-f(\mathbf{x}_i)y_i)) \quad (5)$$

と表現できる。 $f(\mathbf{x}_i)y_i > 0$ のとき \mathbf{x}_i は f によって正しく分類されるため、 $f(\mathbf{x}_i)y_i$ が大きければ大きいほど訓練標本 (\mathbf{x}_i, y_i) を余裕を持って正しく分類できることになる。そのため、 $f(\mathbf{x}_i)y_i$ は訓練標本 (\mathbf{x}_i, y_i) に対するマージン（余裕）とよばれる。式(5)では、マージン $f(\mathbf{x}_i)y_i$ が大きくなるようにパラメータ $\boldsymbol{\theta}$ が学習される。

3.2 最小二乗法

上記のマージン最大化の考え方に基づけば、ロジスティック回帰を様々な学習法に拡張できる。例えば、ロジスティック損失 $\log(1 + \exp(-f(\mathbf{x}_i)y_i))$ の代わりに二乗損失 $(1 - f(\mathbf{x}_i)y_i)^2$ を用いれば、マージンを1に近づけるようにパラメータ $\boldsymbol{\theta}$ が学習される：

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^n (1 - f(\mathbf{x}_i)y_i)^2 \quad (6)$$

ここで、 $y_i = \pm 1$ より

$$(1 - f(\mathbf{x}_i)y_i)^2 = y_i^2 \left(\frac{1}{y_i} - f(\mathbf{x}_i) \right)^2 = (y_i - f(\mathbf{x}_i))^2$$

が成り立つことから、式(6)は訓練出力標本 $\{y_i\}_{i=1}^n$ とモデルの出力 $\{f(\mathbf{x}_i)\}_{i=1}^n$ の二乗誤差和が最小になるようにパラメータ $\boldsymbol{\theta}$ を学習する最小二乗法と等価であることがわかる。最小二乗法は適当な条件のもとで、生成的アプローチで紹介したフィッシャーの判別分析と本質的に一致する。

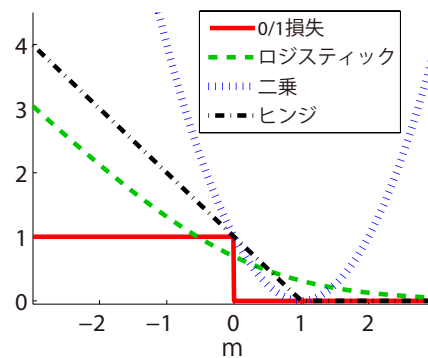


図 1: 損失関数

最小二乗法ではモデルの出力 $\{f(\mathbf{x}_i)\}_{i=1}^n$ を訓練出力標本 $\{y_i\}_{i=1}^n$ に適合させるため、 $\{y_i\}_{i=1}^n$ が雑音を含むときは過適合する傾向がある。そこで、式 (6) に l_2 正則化項 $\|\boldsymbol{\theta}\|^2$ を加えた l_2 正則化最小二乗法

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^n (1 - f(\mathbf{x}_i)y_i)^2 + \lambda \|\boldsymbol{\theta}\|^2$$

がよく用いられる。ただし、 $\lambda \geq 0$ は正則化の強さを調整する正則化パラメータであり、 $\mathcal{X}_y^{(t)}$ の平均対数尤度の代わりに $\mathcal{X}_y^{(t)}$ の誤分類率を用いた交差確認法によって決定できる。 l_2 正則化最小二乗法の解 $\hat{\boldsymbol{\theta}}$ は、解析的に

$$\hat{\boldsymbol{\theta}} = (\mathbf{K}^2 + \lambda \mathbf{I})^{-1} \mathbf{K}(y_1, \dots, y_n)^\top$$

と求められる。ただし、 \mathbf{K} は (i, j) 要素が $K(\mathbf{x}_i, \mathbf{x}_j)$ のカーネル行列であり、 \mathbf{I} は単位行列を表す。「正則化」という名称は、 \mathbf{K}^2 が特異行列の場合でも、 $\lambda \mathbf{I}$ を加えて正則行列にすることに由来している。

l_2 正則化項 $\|\boldsymbol{\theta}\|^2 = \sum_{j=1}^n \theta_j^2$ の代わりに l_1 正則化項 $\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^n |\theta_j|$ を式 (6) に加えると、解 $\hat{\boldsymbol{\theta}}$ がスパース、すなわち、多くの要素がゼロになることが知られている。

3.3 0/1 損失

式 (5) と式 (6) をマージン m の関数と見たものを損失関数とよぶ (図 1) :

$$\text{ロジスティック損失} : \log(1 + \exp(-m))$$

$$\text{二乗損失} : (1 - m)^2$$

パターン認識において、最も自然な損失関数は 0/1 損失であろう (図 1) :

$$\text{0/1 損失} : I(m \leq 0)$$

0/1 損失は、マージンが正のときは0を取り、それ以外のときは1を取る。マージンが正のときにその標本は正しく分類されることから、0/1 損失は誤分類標本数を表す。

0/1 損失は原点で微分不可能で、それ以外の点では傾きを持たない関数であり、この最小化は *NP* 困難であることが知られている。つまり 0/1 損失の最小化は、 n 個の訓練標本の正負のクラスへの 2^n 通りの分類可能性をしらみつぶしに調べるよりも良いアルゴリズムが今のところ知られていない。

従って、しらみつぶし探索が可能なほど訓練標本数 n が小さいとき以外は、0/1 損失を最小にするパラメータ値を求めることが事実上できない。そこで実用上は、ロジスティック損失や二乗損失などの代理損失を近似として用いる。しかし、二乗損失を 0/1 損失の近似としてみたとき、 $m > 1$ で損失が増加するのが不自然である (図 1)。

3.4 サポートベクトルマシン

損失最小化の観点から、二乗損失の代わりによく用いられるのがヒンジ損失である。

$$\text{ヒンジ損失} : \max(0, 1 - m)$$

図 1 に示したように、「ヒンジ損失」という名称は、ちょうつがい (ヒンジ) を 135 度開いた形状をしていることに由来している。

ヒンジ損失に一般化した l_2 正則化項 $\theta^\top \mathbf{K} \theta$ を加えた規準を最小にするパターン認識法

$$\min_{\theta} \sum_{i=1}^n \max(0, 1 - f(\mathbf{x}_i) y_i) + \lambda \theta^\top \mathbf{K} \theta \quad (7)$$

は、サポートベクトルマシンよばれている。歴史的には、サポートベクトルマシンはソフトマージン最大化やカーネルトリックとよばれる技法を組合せて導出されたが、上述したように最小二乗法の自然な拡張としても導出できる。

技術的な詳細は省略するが、最適化問題 (7) のラグランジュ双対問題を考えると、双対変数 $\alpha_i = y_i \theta_i$ に関する最適化問題

$$\max_{0 \leq \alpha \leq \frac{1}{\lambda} \mathbf{1}} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

が得られる。ただし、 $\mathbf{0}$ と $\mathbf{1}$ はそれぞれ 0 と 1 だけからなるベクトルを表し、ベクトルに対する不等号は要素毎の大小を表す。これは、二次計画とよばれる最適化問題であり、標準的な最適化ソフトウェアで解を求められる。また、この双対問題の解はスパースになることが知られている。

式 (4) で示したカーネルモデルの定義より、 n 次元の双対変数 α の各要素は各訓練標本に対応している。「サポートベクトル」という名称は、 $\theta_i = \alpha_i y_i$ という関係より、非零の双

対変数 α_i に対応する訓練入力標本ベクトル \mathbf{x}_i だけが学習結果を支えて（サポートして）いることに由来している。

また、サポートベクトルマシンは各クラスの事後確率の差の符号

$$\text{sign}\left(p(y = +1|\mathbf{x}) - p(y = -1|\mathbf{x})\right) \quad (8)$$

を推定しているとも解釈できる。式 (8) がまさにパターン \mathbf{x} をクラス $y = \pm 1$ に分類するために直接的に必要な量であり、それを直接推定していることがサポートベクトルマシンのロジスティック回帰に対する優位性だと考えられる。

一方ロジスティック回帰では、パターン \mathbf{x} の認識に対する信頼度が事後確率 $p(y|\mathbf{x})$ として得られるため、認識が難しいパターンを棄却するといった使い方が可能である。

4 機械学習の知識を更に深めるには

現在主流の機械学習技術は確率と統計を基盤としており、MATLAB や R などを用いれば容易に基礎的なデータ解析が行える [2, 3, 4].

一方、最先端の機械学習手法は最適化理論やアルゴリズム論などとも深く関わっているため、初学者が習得するのは必ずしも容易でない。そして、これらのアルゴリズムを用いて実際にデータ解析を行うためには、MapReduce や Hadoop などの分散処理技術、GPU プログラミング、更には、データベースシステムや計算機システムそのものに関する知識が必要となることもあり、敷居が低いとは言えない。また、機械学習技術の応用分野は基礎科学から産業まで非常に多様なため、これらを俯瞰的な視点から学ぶことも困難である。このような状況を踏まえて、発展著しい機械学習技術の数学的な基礎理論、実用的なアルゴリズム、そして、それらの活用法を、入門的な内容から先端的な研究成果までわかりやすく解説した教科書も登場しつつある [6].

本稿が、読者の機械学習への興味を掻き立てるとともに、ビッグデータ時代を渡り歩いていくための一助となることを願う。

参考文献

- [1] 元田浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田昇 (編). パターン認識と機械学習 (上・下). 丸善出版, 2007~2008.
- [2] 金森敬文, 竹之内高志, 村田昇. パターン認識, 共立出版, 2009.
- [3] 杉山将. 統計的機械学習. オーム社, 2009.
- [4] 杉山将. イラストで学ぶ機械学習. 講談社, 2013.
- [5] 杉山将, 井手剛, 神畷敏弘, 栗田多喜夫, 前田英作 (編). 統計的学習の基礎. 共立出版, 2014.
- [6] 杉山将 (編), 機械学習プロフェッショナルシリーズ (全 29 巻). 講談社, 2015.