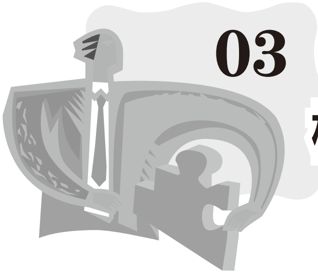
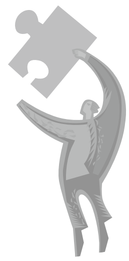


基

03 機械学習のための数学



杉山 将（東京大学） 鈴木大慈（東京工業大学）

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 $\boldsymbol{x}\in\mathbb{R}^d$ をそれが属するクラス $y\in\{1,\ldots,c\}$ 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．たとえば，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 × 16＝256 であり，クラス数は数字の 0 ～ 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 $\boldsymbol{x}$ が従う確率 $p(\boldsymbol{x})$，クラス y が従う確率 $p(y)$，それらの同時確率 $p(\boldsymbol{x},y)$，パターン $\boldsymbol{x}$ がクラス y に属する条件付き確率 $p(y|\boldsymbol{x})$，クラス y に属するパターン $\boldsymbol{x}$ の条件付き確率 $p(\boldsymbol{x}|y)$ を考える☆1．$p(y|\boldsymbol{x})$ はパターン $\boldsymbol{x}$ を見た後でのクラス y の出現確率，$p(y)$ はパターン $\boldsymbol{x}$ 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**と呼ぶ．事後確率 $p(y|\boldsymbol{x})$ が最大となるクラス y にパターン $\boldsymbol{x}$ 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 $p(y|\boldsymbol{x})$ が未知であるため，同時確率 $p(\boldsymbol{x},y)$ に独立に従う n 個の訓練標本 $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$ を用いてパターン認識器を構成することにする．$p(y|\boldsymbol{x})$ を訓練標本から直接推定する方式を**識別的アプローチ** [1), 2)] と呼ぶ．「識別的」という名称は，パターン $\boldsymbol{x}$ をクラス y に識別するために必要な事後確率 $p(y|\boldsymbol{x})$ 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**

$$p(y|\boldsymbol{x})=\frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})}\propto p(\boldsymbol{x}|y)p(y) \qquad (1)$$

を用いれば，$p(y|\boldsymbol{x})$ の y に関する最大化を $p(\boldsymbol{x}|y)p(y)$ の最大化に変換できる．ここで，$\propto$ は比例関係を表す．$p(\boldsymbol{x}|y)p(y)$ を訓練標本から推定することによってパターン認識を行う方式を，**生成的アプローチ** [3), 4)] と呼ぶ．「生成的」という名称は，データの生成確率 $p(\boldsymbol{x}|y)p(y)=p(\boldsymbol{x}, y)$ を推定することに由来している．

生成的アプローチ

生成的アプローチでは，$p(\boldsymbol{x}|y)$ と $p(y)$ を訓練標本 $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$ から推定する．$p(y)$ は，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，$p(\boldsymbol{x}|y)$ の推定にはさまざまな方法が用いられる．

❖最尤推定法

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

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

を最大にするようにパラメータ $\boldsymbol{\theta}_y$ を決定する．すなわち，$p(\boldsymbol{x}|y)$ の推定量 $\hat{p}(\boldsymbol{x}|y)$ は

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

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 $\mathcal{X}_y=\{\boldsymbol{x}_i\}_{i:y_i=y}$ のもとで，最も尤

☆1 正確には，$p(x)$ 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」と呼ぶことにする．

もらしいパラメータ値を求めることに由来している．

たとえば，パラメトリックモデル $q(\boldsymbol{x}|\boldsymbol{\theta}_y)$ 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**

$$q_{\mathrm{G}}\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_y,\Sigma_y\right)=\left(2\pi\right)^{-\frac{d}{2}}\left|\Sigma_y\right|^{-\frac{1}{2}}$$
$$\times\exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_y\right)^{\top}\Sigma_y^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_y\right)\right) \tag{2}$$

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 $\sum_{i:y_i=y}\log q_{\mathrm{G}}\left(\boldsymbol{x}_i\middle|\boldsymbol{\mu}_y,\Sigma_y\right)$ の微分をゼロとおけば，最尤推定量 $\hat{\boldsymbol{\mu}}_y$, $\hat{\boldsymbol{\Sigma}}_y$ は，

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

と解析的に求められる．

全クラス $y=1,\ldots,c$ 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

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

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

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

と近似できる．ここで，C は y に依存しない定数である．したがって，クラス間の分離境界は $\boldsymbol{x}$ の一次形式となる．これを**フィッシャーの判別分析**と呼ぶ．

❖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**

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

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 $q(\boldsymbol{x}|\boldsymbol{\theta}_y)$ と事前確率 $p(\boldsymbol{\theta}_y)$ に依存し，組合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 $q(\boldsymbol{x}|\boldsymbol{\theta}_y)$ 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．たとえば，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**

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

が共役な事前確率である．ここで，$\boldsymbol{m}$ は d 次元ベクトル，κ は正のスカラー，ν は正の整数，$\boldsymbol{S}$ 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

$$\boldsymbol{m}\leftarrow\frac{\kappa\boldsymbol{m}+n_y\hat{\boldsymbol{\mu}}_y}{\kappa+n_y},\ \kappa\leftarrow\kappa+n_y,\ \nu\leftarrow\nu+n_y,$$
$$\boldsymbol{S}\leftarrow\boldsymbol{S}+n_y\widehat{\Sigma}_y+\frac{\kappa n_y}{\kappa+n_y}\left(\hat{\boldsymbol{\mu}}_y-\boldsymbol{m}\right)\left(\hat{\boldsymbol{\mu}}_y-\boldsymbol{m}\right)^{\top}$$

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 $q(\boldsymbol{x}|\boldsymbol{\theta}_y)$ の事後確率に関する期待値

$$\bar{p}\left(\boldsymbol{x}\middle|y\right)=\int q\left(\boldsymbol{x}\middle|\boldsymbol{\theta}_y\right)p\left(\boldsymbol{\theta}_y\middle|\mathcal{X}_y\right)\mathrm{d}\boldsymbol{\theta}_y \tag{3}$$

を $p(\boldsymbol{x}|y)$ の推定結果とする．これを**ベイズ予測分布**と呼ぶ．最尤推定では，モデル $q(\boldsymbol{x}|\boldsymbol{\theta}_y)$ 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 $q(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_y)$ を求めた．したがって，たとえば式 (2) で示したガウスモデルを用いれば，推定量 $q(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_y)$ もガウス分布である．一方，ベイズ予測分布はモデル $q(\boldsymbol{x}|\boldsymbol{\theta}_y)$ を平均化するため，たとえ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**と呼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,

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

と $p(\boldsymbol{x}|y)$ 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分かる．そのため，最大事後確率推定法は**罰則付き最尤推定法**と呼ばれることもある．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．

❖カーネル密度推定法

最尤推定法やベイズ推定法では，$p(\boldsymbol{x}|y)$ の推定にパラメトリックモデル $q(\boldsymbol{x}|\boldsymbol{\theta}_y)$ 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 $p(\boldsymbol{x}|y)$ 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,

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

によって $p(\boldsymbol{x}|y)$ を推定する．ここで，$K(\boldsymbol{x}, \boldsymbol{x}')$ は**カーネル**と呼ばれる 2 変数関数であり，**ガウスカーネル**

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

がよく用いられる．$h>0$ は**バンド幅**と呼ばれ，ガウスカ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，たとえば以下の**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 $\mathcal{X}_y = \{\boldsymbol{x}_i\}_{i:y_i=y}$ 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 $\hat{p}_h^{(t)}(\boldsymbol{x}|y)$ を求め，$\mathcal{X}_y^{(t)}$ での平均対数尤度を求める．
$$L_h^{(t)} = \frac{1}{|\mathcal{X}_y^{(t)}|} \sum_{\boldsymbol{x}^{(t)} \in \mathcal{X}_y^{(t)}} \log \hat{p}_h^{(t)}(\boldsymbol{x}^{(t)}|y)$$
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. これをさまざまなバンド幅 h に対して実行し，L_h を最大にするものを選択する．

❖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 $\mathcal{X}_y = \{\boldsymbol{x}_i\}_{i:y_i=y}$ の中で注目点 $\boldsymbol{x}$ から k 番目に近い点 $\boldsymbol{x}^{(k)}$ への距離 $r_y(k) = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|$ を用いて,

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

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

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-x}\, \mathrm{d}u$$

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 $p(y|\boldsymbol{x})$ 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：

$$\underset{y}{\operatorname{argmax}}\ \hat{p}(\boldsymbol{x}|y)\hat{p}(y) = \underset{y}{\operatorname{argmax}} \frac{\Gamma\left(\frac{\mathrm{d}}{2}+1\right)}{n_y \pi^{\frac{d}{2}} r_y(1)^d} \frac{n_y}{n}$$
$$= \underset{y}{\operatorname{argmax}}\ r_y(1)^{-d} = \underset{y}{\operatorname{argmin}}\ r_y(1)$$

すなわち，全クラスの訓練パターン $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ の中で注目点 $\boldsymbol{x}$ に最も近いパターンが属するクラス y を，$\boldsymbol{x}$ が属するクラスと予測する．これを**最近傍識別器**と呼ぶ．これを一般化して，注目点 $\boldsymbol{x}$ 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**と呼ぶ．近傍数 k の値は，たとえば上述した**交差確認法**によって決定できる．

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 $p(y|\boldsymbol{x})$ を直接推定する．

❖ロジスティック回帰

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

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

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 $\left\{\theta_j^{(y)}\right\}_{y=1,j=1}^{c,n}$ 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，たとえば**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：

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

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 個の訓練標本すべ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 $y=\pm1$ とおき，パラメータ $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_n)^{\top}$ を持つ**カーネルモデル**

$$f\left(\boldsymbol{x}\right)=\sum_{j=1}^{n}\theta_jK(\boldsymbol{x},\boldsymbol{x}_j) \tag{4}$$

を用いて，ロジスティックモデルを

$$q_{\mathrm{L}}\left(y=+1\middle|\boldsymbol{x}\right)=\frac{\exp\left(f\left(\boldsymbol{x}\right)\right)}{\exp\left(f\left(\boldsymbol{x}\right)\right)+\exp\left(-f\left(\boldsymbol{x}\right)\right)}$$
$$q_{\mathrm{L}}\left(y=-1\middle|\boldsymbol{x}\right)=1-q_{\mathrm{L}}\left(y=+1\middle|\boldsymbol{x}\right)$$

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

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

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

❖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さまざまな学習法に拡張できる．たとえば，ロジスティック損失 $\log\left(1+\exp\left(-f\left(\boldsymbol{x}_i\right)y_i\right)\right)$ の代わりに二乗損失 $\left(1-f\left(\boldsymbol{x}_i\right)y_i\right)^2$ 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：

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

ここで，$y_i=\pm1$ より

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

が成り立つことから，式 (6) は訓練出力標本 $\left\{y_i\right\}_{i=1}^{n}$ とモデルの出力 $\left\{f\left(\boldsymbol{x}_i\right)\right\}_{i=1}^{n}$ 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分かる．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．

最小二乗法ではモデルの出力$\left\{f\left(\boldsymbol{x}_i\right)\right\}_{i=1}^n$を訓練出力標本$\left\{y_i\right\}_{i=1}^n$に適合させるため，$\left\{y_i\right\}_{i=1}^n$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$\ell_2$**正則化最小二乗法**

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

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

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

と求められる．ただし，$\boldsymbol{K}$は(i, j)要素が$K(\boldsymbol{x}_i, \boldsymbol{x}_j)$の**カーネル行列**であり，$\boldsymbol{I}$は単位行列を表す．「正則化」という名称は，$\boldsymbol{K}^2$が特異行列の場合でも，$\lambda\boldsymbol{I}$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$\|\theta\|^2=\sum_{j=1}^n\theta_j^2$の代わりに$\ell_1$**正則化項**$\|\theta\|_1=\sum_{j=1}^n|\theta_j|$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．

❖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**と呼ぶ（**図-1**）：

ロジスティック損失：$\log(1+\exp(-m))$

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）：

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．

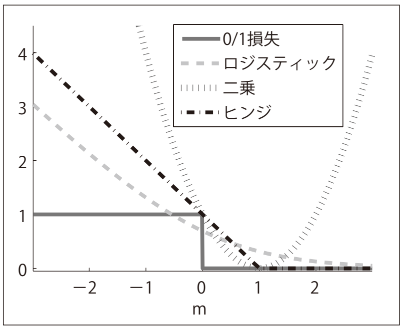


図-1　損失関数

したがって，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）．

❖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．

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**$\boldsymbol{\theta}^{\top}\boldsymbol{K}\boldsymbol{\theta}$を加えた規準を最小にするパターン認識法

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

は，**サポートベクトルマシン**と呼ばれている．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**と呼ばれる技法を組み合わせて導出されたが，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

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

が得られる．ただし，**0** と **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ごとの大小を表す．これは，**二次計画**と呼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 $\boldsymbol{\alpha}$ 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 $\boldsymbol{x}_i$ 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

$$\mathrm{sign}\Big(p\big(y=+1|\boldsymbol{x}\big)-p\big(y=-1|\boldsymbol{x}\big)\Big) \qquad (8)$$

を推定しているとも解釈できる．式 (8) こそがまさにパターン $\boldsymbol{x}$ をクラス $y=\pm1$ 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 $\boldsymbol{x}$ の認識に対する**信頼度**が事後確率 $p(y|\boldsymbol{x})$ 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．

機械学習の知識をさらに深めるには

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), 4), 5)]．

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かかわっているため，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，さらには，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分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) 杉山　将，井手　剛，神嶌敏弘，栗田多喜夫，前田英作 編：統計的学習の基礎，共立出版 (2014).
2) 杉山　将：イラストで学ぶ機械学習，講談社 (2013).
3) 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).
4) 杉山　将：統計的機械学習，オーム社 (2009).
5) 金森敬文，竹之内高志，村田　昇：パターン認識，共立出版 (2009).
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)，http://urx2.nu/gc1Q

(2015 年 1 月 8 日受付)

杉山　将（正会員） ■ sugi@k.u-tokyo.ac.jp

2001 年に東京工業大学より博士（工学）の学位を取得．現在，東京大学教授．機械学習とデータマイニングのアルゴリズムの開発，および，その応用研究に従事．

鈴木大慈 ■ s-taiji@is.titech.ac.jp

2009 年に東京大学より博士（情報理工学）の学位を取得．現在，東京工業大学准教授．専門は統計学と機械学習，特に統計的学習理論および最適化手法の研究に従事．