第 4 章 ベイズ判別問題とパターン認識

4.4 ミニマックス決定
4.5 パターン認識と分類
4.6 ベイズ決定による判別（分類）
4.7 判別分析

05T4078Y **茂木哲矢**

2008.10.03

Outline

トレード・オフ

伝送の確率

- 「イエス」$\rightarrow$ 「ノー」の誤りが起こりやすい
- $e(1|0)$ は減少すると，$e(0|1)$ は増加する
- $e(0|1)$ は減少すると，$e(1|0)$ は増加する
- $e(1|0)$ と $e(0|1)$ はトレード・オフの関係

表 4.5: 伝送の確率

送信 $\rightarrow$ 受信	確率 (%)
「イエス」$\rightarrow$「イエス」	93.970
「イエス」$\rightarrow$「ノー」	6.030
「ノー」$\rightarrow$「ノー」	97.220
「ノー」$\rightarrow$「イエス」	2.780

ミニマックス決定

- 誤りの確率の大きいほう $Max\{e(0|1), e(1|0)\}$ を最小化する
- $e(0|1) = e(1|0)$ となるようにする
- 2 つのくさび形が等しくなる決定境界で，判断の方式 $x \gtreqless 0.500$ のとき
- 計算の際に，事前確率，損失の情報がいらない

パターン認識と例

パターン認識

- パターンを概念に分類要約して，それに対応付けること
- 数理的問題としてはかなり難しい
- 普通の意味で感覚による認識は，すべてパターン認識

パターン認識の例

- 光学的文字読取装置
- マークシートや郵便番号の読取装置
- スキャナー付属の文字読み取り機能

パターン認識の難しさ

- 2 次元，3 次元のパターンでは，どこから計算するか決めることが難しい
- コンピュータによる計算は，ディジタル的なものには適しているが，アナログ的なものには適さない

→ 解を見つける方法の 1 つとして，統計的決定論がある

特徴抽出と判別・分類

特徴抽出

- パターンとは図形そのものではない
- パターンは様々な特徴を持つ（花の形では，花弁の大きさ，がく片の大きさ，etc.）
- そのものとわかるための主要な特徴を選出する

$$\Psi\,(\text{無限次元}) \to z\ (p\ \text{次元})$$

Ψ はパターン，z は特徴ベクトル

- 特徴抽出 $=$ 次元を低下させる

判別・分類

特徴ベクトルから，カテゴリーのどれかに分類する数理的手続きのこと

ベイズ決定 事前分布

z が分からないとき Π に分類する問題

損失 $0-1$ 型単純損失

事前分布 $w_k \ (k=1,\ 2,\ \cdots,\ K)$

Π_k に分類する事前期待損失

$$w_1\cdot 1+\cdots+w_{k-1}\cdot 1+w_k\cdot 0+w_{k+1}\cdot 1+\cdots+w_K\cdot 1 \quad (4.6.3)$$

Π_k に分類する条件

$$1-w_k \leqq 1-w_l \quad (l=1,2,\cdots,K) \quad (4.6.4)$$

ベイズ決定 事後分布

z が分かったとき Π に分類する問題

事後確率最大の Π へ分類する

Π_k の事後確率

$$w'_k = \frac{w_k f_k(z)}{w_1 f_1(z) + \cdots + w_K f_K(z)} \tag{4.6.5}$$

Π_k に分類する条件

$$w_k f_k(z) = \max\{w_1 f_1(z), \cdots, w_K f_K(z)\} \tag{4.6.6}$$

Π_k に分類する条件（尤度比の形）

$$\frac{f_1(z)}{f_k(z)} \leqq \frac{w_k}{w_1}, \frac{f_2(z)}{f_k(z)} \leqq \frac{w_k}{w_2}, \cdots, \frac{f_K(z)}{f_k(z)} \leqq \frac{w_k}{w_K} \tag{4.6.7}$$

多次元正規分布 I

$\boldsymbol{z} = (x_1, x_2, \cdots, x_p)'$ の分布 $f(\boldsymbol{z})$

平均

$$\mu_i = E(x_i) \quad (i = 1, 2, \cdots, p) \tag{4.6.9}$$

共分散

$$\sigma_{ij} = E((x_i - \mu_i)(x_j - \mu_j)) \quad (i, j = 1, 2, \cdots, p) \tag{4.6.10}$$

p 次元正規分布 $N_p(\theta, \Sigma)$

$$f(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\theta})\},$$
$$\boldsymbol{\theta} = (\mu_1, \mu_2, \cdots, \mu_p), \quad \boldsymbol{\Sigma} = \{\sigma_{ij}\} \tag{4.6.11}$$

多次元正規分布 II

$z = p$ 次元ベクトルの場合の尤度比

$$\log \frac{f_z(\boldsymbol{z})}{f_k(\boldsymbol{z})} = (\boldsymbol{\theta}_l - \boldsymbol{\theta}_k)' \boldsymbol{\Sigma}^{-1} \boldsymbol{z} - \frac{1}{2} (\boldsymbol{\theta}_l + \boldsymbol{\theta}_k)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_l - \boldsymbol{\theta}_k) \equiv F_{lk}(\boldsymbol{z})$$

$$F_{lk}(\boldsymbol{z}) \leqq \log \left(\frac{w_k}{w_l} \right) \quad (l = 1, 2, \cdots, K) \tag{4.6.13}$$

- $\boldsymbol{z}$ に対する F_{lk} の値で分類される
- F_{lk} は $\boldsymbol{z}$ に対して線型なので線型判別関数とよばれる

$$F_{kl}(\boldsymbol{z}) = -F_{lk}(\boldsymbol{z}) \tag{4.6.14}$$

ベイズ判別領域

$$R_k = \{ \boldsymbol{z} | F_{kl} \geqq \log \left(\frac{w_l}{w_k} \right) \quad (l = 1, 2, \cdots, K) \} \tag{4.6.15}$$

判別分析 I

判別分析

- 多変量解析の方法の 1 つ
- 固有値問題に帰着する
- 次元を落とすことによって情報の圧縮ができる
  ($=$ 特徴抽出に用いることができる)

判別分析の問題

群（母集団，カテゴリー） K 個　**ケース（個体，サンプル）の数**

群に対する変数 $x_i(i = 1, 2, \cdots, p)$　$n_k(k = 1, 2, \cdots, K)$

線型関数

$$L = l_1 x_1 + l_2 x_2 + \cdots + l_p x_p \tag{4.7.1}$$

L の “差が付く” ように，$l_1, l_2, \cdots, l_p$ を選ぶ

分散・共分散行列

x_j の平均

$$\bar{x}_j^{(s)} = \frac{\sum_i^{(s)} x_{ij}}{n_s}, \quad \bar{x}_j = \frac{\sum_i x_{ij}}{n} \tag{4.7.2}$$

変数間の情報をとらえるために，分散・共分散行列を用いる．

$$\begin{aligned} \textbf{群間：} b_{jl} &= \frac{\sum_{s=1}^{K} n_s(\bar{x}_j^{(s)} - \bar{x}_j)(\bar{x}_l^{(s)} - \bar{x}_l)}{n-1} \\ \textbf{群内：} w_{jl} &= \frac{\sum_{s=1}^{K} \sum_i^{(s)} (x_{ij}^{(s)} - \bar{x}_j^{(s)})(x_{il}^{(s)} - \bar{x}_l^{(s)})}{n-1} \\ \textbf{全体：} t_{jl} &= \frac{\sum_i (x_{ij} - \bar{x}_j)(x_{il} - \sum_{\substack{l \\ i}}^{(s)})}{n-1} \end{aligned} \tag{4.7.4}$$

$$\boldsymbol{B} = \{b_{ij}\}, \quad \boldsymbol{W} = \{w_{ij}\}, \quad \boldsymbol{T} = \{t_{ij}\} \tag{4.7.5}$$

$$t_{jl} = b_{jl} + w_{jl}, \quad \boldsymbol{T} = \boldsymbol{B} + \boldsymbol{W} \tag{4.7.6}$$

判別スコア I

判別スコア

判別スコア $L = l_1x_1 + l_2x_2 + \cdots + l_px_p$

判別スコアは，群間分散 $\gg$ 群内分散のような値であることが望ましい

$\boldsymbol{l} = \{l_1, l_2, \cdots, l_p\}'$ **とする**

群間分散

$$\boldsymbol{l}'\boldsymbol{B}\boldsymbol{l} = \sum_i \sum_j b_{ij} l_i l_j \tag{4.7.8}$$

群内分散

$$\boldsymbol{l}'\boldsymbol{W}\boldsymbol{l} = \sum_i \sum_j w_{ij} l_i l_j \tag{4.7.9}$$

レイリー商

$$Q \equiv \frac{\boldsymbol{l}'\boldsymbol{B}\boldsymbol{l}}{\boldsymbol{l}'\boldsymbol{W}\boldsymbol{l}} \Rightarrow \textbf{最大} \tag{4.7.10}$$

判別スコア II

固有値問題

群の間で "差が付く" l の選び方 $=$ $$\boldsymbol{Bl = \lambda W l} \tag{4.7.11}$$

に帰着する

最大固有値

$$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p \tag{4.7.12}$$

固有ベクトル $l^{(1)}$ の成分 $l_1^{(1)}, l_2^{(1)}, \cdots, l_p^{(1)}$

判別スコア（判別関数）

$$L^{(1)} = l_1^{(1)} x_1 + l_2^{(1)} x_2 + \cdots l_p^{(1)} x_p \tag{4.7.13}$$

まとめ

- 誤りがトレード・オフの関係にあるとき決定境界はミニマックス決定で決める
- パターン認識の解を見つける方法には
  - ベイズ決定
  - 判別分析

  がある