線形判別ルールと最大尤度比ルールの優劣について

若木宏文

99.5.11, 数学教室談話会

分散共分散行列が等しい二つの多変量正規母集団、$\Pi_1 : N_p(\mu_1, \Sigma), \Pi_2 : N_p(\mu_2, \Sigma)$ への判別問題を考える。μ_1, μ_2, Σ は未知母数であり、これら (母集団) に関する情報として、母集団 $\Pi_i (i = 1, 2)$ からの標本 :$X_1^{(i)}, X_2^{(i)}, \cdots, X_{N_i}^{(i)}$ が利用できるとする。X を Π_1 または Π_2 から観測される変数とする。判別問題は、X の値と、各母集団からの標本 (参照標本と呼ぶ) の値に基づいて X が Π_1 と Π_2 のどちらから観測されたものかを判定する問題である。参照標本の値が与えられたとき、判別ルールを構成する問題は、X の取り得る値の全体からなる集合、すなわち、p-次元ユークリッド空間 R^p を共通部分を持たない 2 つの部分集合 R_1 と R_2 に分割する問題として表される。X の値が R_1 に含まれるとき、X は Π_1 から観測されたものと判定することで、判別ルールと R^p の部分集合 R_1 とは 1 対 1 に対応する。$R_2 = R^p \setminus R_1$ であるから、判別ルールを、対応する部分集合 R_1 によって表す。

判別ルール R の良し悪しの基準の一つは、適当な正の定数 c_1, c_2 によって、$r(R) = c_1 \Pr\{X \in R^c | \Pi_1\} + c_2 \Pr\{X \in R | \Pi_2\}$ と表され、$r(R)$ が小さいほど、R は望ましい判別ルールとされる。μ_1, μ_2, Σ の値がわかっているときには、$r(R)$ を最小にする判別ルールは、$R_o = \{x | x'\Sigma^{-1}(\mu_1 - \mu_2) > \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) + \log(c_2/c_1)\}$ によって与えられる。

未知母数 μ_1, μ_2, Σ の値が未知の場合の判別ルールとして代表的なものに (標本) 線形判別ルールと最大尤度比ルールがある。線形判別ルールは、W ルールとも呼ばれ、未知母数 μ_1, μ_2, Σ の値を、それぞれの不偏推定量、$\overline{X}_1, \overline{X}_2, S$ で置き換えたものである。ここで、$\overline{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_j^{(i)}, (i = 1, 2), S = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{N_i}(X_j^{(i)} - \overline{X}_i)(X_j^{(i)} - \overline{X}_i)', n = N_1 + N_2 - 2$ である。最大尤度比ルールは Z ルールとも呼ばれ、判別問題を 2 つの仮説

$$
\begin{aligned}
H_1 &: X_1^{(1)}, \cdots, X_{N_1}^{(1)}, X \quad \sim N_p(\mu_1, \Sigma), \quad X_1^{(2)}, \cdots, X_{N_2}^{(2)} \quad \sim N_p(\mu_2, \Sigma) \\
H_2 &: X_1^{(1)}, \cdots, X_{N_1}^{(1)} \quad \sim N_p(\mu_1, \Sigma), \quad X_1^{(2)}, \cdots, X_{N_2}^{(2)}, X \quad \sim N_p(\mu_2, \Sigma)
\end{aligned}
$$

の仮説検定と考えることによって得られる。Z ルールは $R_Z = \{x | \frac{N_1}{N_1+1} d(x, \overline{X}_1; S) - \frac{N_2}{N_2+1} d(x, \overline{X}_2; S) < c\}$ と表される。ここで、$d(x, y; Z) = (x - y)'Z^{-1}(x - y), c$ は $r(R_Z)$ が小さくなるように c_1, c_2 の比に応じて決められる。

線形判別ルールと最大尤度比ルールの優劣を調べるため、$N_1 \longrightarrow \infty, N_2 \longrightarrow \infty, \frac{N_1}{N_2} \longrightarrow \kappa (\kappa > 0)$ のときの各ルールの $r(R)$ の漸近展開を求める。線形判別ルールは、$R_W = \{x | d(x, \overline{X}_1; S) - d(x, \overline{X}_2; S) < 2\log(c_1/c_2)\}$ と表されるので、単に 2 つの判別ルールの優劣を調べるだけでなく、R_W, R_Z を含む判別ルールのクラス, $R(a, b) = \{x | (1 + a)d(x, \overline{X}_1; S) - (1 - a)d(x, \overline{X}_2; S) < b\}$ を考え、$r(R(a, b))$ の漸近展開の $O(n^{-2})$ までの項を最小にするような係数 a, b の値を求めることを考えた。

$r(R(a, b))$ は a, b だけでなく、Π_1, Π_2 のマハラノビス距離 $\Delta = d(\mu_1, \mu_2; \Sigma)^{1/2}$ の値にも依存しており、Δ に関して一様に $r(R(a, b))$ を最小にすることが望まれる。残念ながら、一様に最小にする a, b は存在しないが、任意に b を固定したとき、Δ に関して一様に $r(R(a, b))$ を最小にする a を求めることができた。$c_1 = c_2$ の場合には最適な a は、最大尤度比ルールに対応するものと $O(n^{-1})$ まで一致しており、漸近展開における 2 次のオーダーまでは、最大尤度比ルールは、線形判別ルールより優れていることが示された。