高次元データにおけるバイアス補正非線形 SVM について

筑波大学・数理物質科学　　中山　優吾
筑波大学・数理物質系　　矢田　和善
筑波大学・数理物質系　　青嶋　　誠

本講演では，高次元小標本データに対する判別分析を考える．母集団が 2 個あると想定し，各母集団 π_i $(i=1,2)$ は平均に p 次ベクトル $\boldsymbol{\mu}_i$，共分散行列に p 次正定値対称行列 $\boldsymbol{\Sigma}_i$ $(>\boldsymbol{O})$ をもつと仮定する．各母集団 π_i から n_i (≥ 2) 個の学習データ $\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}$ を無作為に抽出する．$p > n_1+n_2$ と仮定する．判別対象の p 次元データを $\boldsymbol{x}_0$ とし，$\boldsymbol{x}_0 \in \pi_1$ もしくは $\boldsymbol{x}_0 \in \pi_2$ を仮定し，$\boldsymbol{x}_0 \in \pi_1$ のときに判別対象を誤判別する確率を $e(1)$ とし，$e(2)$ も同様の表記とする．$\Delta = \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2$ とおく．Vapnik 等が考案したサポートベクターマシン（SVM）は高次元データ解析において疎な解を与え，汎化性能が良いことが知られているが，SVM の精度保証については理論的な研究が乏しい．これに対し，Nakayama et al. [1] は線形 SVM の漸近的性質を高次元小標本の枠組みで理論的に研究し，線形 SVM について，仮定 (A-i): $\limsup_{p\to\infty}|\mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1-\mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2|/\Delta<1$ と適当な正則条件のもと，

$$e(i)\to 0 \text{ as } p\to\infty \text{ for } i=1,2 \tag{1}$$

なる一致性が得られることを証明した．しかしながら，n_1 と n_2（もしくは，$\mathrm{tr}(\boldsymbol{\Sigma}_1)$ と $\mathrm{tr}(\boldsymbol{\Sigma}_2)$）が不均等の場合，(A-i) は仮定できない．それに対し，Nakayama et al. [1] ではバイアス補正線形 SVM を提案し，仮定 (A-i) なしで一致性が与えられることを示した．

本講演では，ガウスカーネルを用いた非線形 SVM（GSVM）を考え，高次元小標本の枠組みでその漸近的性質を導出する．いま，$\gamma>0$ に対し，$\beta_i=\exp\{-2\mathrm{tr}(\boldsymbol{\Sigma}_i)/\gamma\}$，$i=1,2$，$\beta_3=\exp[-\{\mathrm{tr}(\boldsymbol{\Sigma}_1)+\mathrm{tr}(\boldsymbol{\Sigma}_2)+\Delta\}/\gamma]$ とし，$\Delta_*=\beta_1+\beta_2-2\beta_3$ とおく．ここで，$\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$ もしくは $\mathrm{tr}(\boldsymbol{\Sigma}_1)\neq\mathrm{tr}(\boldsymbol{\Sigma}_2)$ のとき，$\Delta_*>0$ となることに注意する．そのとき，次が成り立つ．

定理 1．　仮定 (A-ii): $\limsup_{p\to\infty}|(1-\beta_1)/n_1-(1-\beta_2)/n_2|/\Delta_*<1$ と適当な正則条件のもと，GSVM について (1) が成り立つ．

GSVM は $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ のもとでも一致性をもつことに注意する．さらに，(A-ii) は GSVM のバイアス項に関する仮定であることに注意し，(A-ii) が仮定できない場合は以下の不一致性が成り立つ．

系 1．　GSVM の判別関数に対し，適当な正則条件のもと，$p\to\infty$ で以下が成り立つ．

$$e(1)\to 1 \ \text{ and } \ e(2)\to 0 \text{ if } \ \liminf_{p\to\infty}\frac{(1-\beta_1)/n_1-(1-\beta_2)/n_2}{\Delta_*}>1; \quad \text{and}$$
$$e(1)\to 0 \ \text{ and } \ e(2)\to 1 \text{ if } \ \limsup_{p\to\infty}\frac{(1-\beta_1)/n_1-(1-\beta_2)/n_2}{\Delta_*}<-1.$$

それゆえ，バイアス項に依存して，高次元における GSVM の判別精度が極端に悪くなり得る．

当日は，そのバイアス項を補正したバイアス補正 GSVM を提案し，仮定 (A-ii) なしで一致性が与えられることを示し，その判別性能を数値実験と実データ解析を用いて検証する．さらに，一般のカーネルにも拡張し，新たなバイアス補正非線形 SVM も提案する．

[1] Nakayama, Y., Yata, K., Aoshima, M. (2017). Support vector machine and its bias correction in high-dimension, low-sample-size settings. *J. Stat. Plan. Infer.*, in press.