判別分析におけるクロスバリデーションの信頼性

中西　寛子　　大学統計教員育成センター　特任教授

【はじめに】

・クロスバリデーション（Cross-Validation, CV）

判別分析に応用するCVの基本的な操作について述べる．サイズnのデータをおおよそ等しく K個の集団に分ける．その中の１つの集団を残し，K-1個の集団を使って判別ルールを構築する．残した集団に含まれるデータをそのルールに応用し，ルールの良さを比較する評価基準（誤判別率）を計算する．これを順次，K回繰り返す．K回の評価基準の和や平均などを考え，ルールに対する最終評価を行う．K個の集団に分けることを明確にするときにはK-fold CVという．K=nの場合，残しておく観測値は1つであり，n回検証を繰り返すことになる．これを一つ取って置き法（Leave-One-Out, LOO）とよび，一般のK-fold CVと区別する．

・本研究の目的

本研究の目的はCVを用いたときの実際の誤判別率の信頼区間を理論と数値実験により推定することである．紙面の関係上，ここではLOOの実際の誤判別率の信頼区間を求めるまでの手順をLachenbruch [4] の研究に従い示す．いくつかの数値実験の結果より現実の問題点について最後に言及する．一般のK-fold CVについてもほぼ同様に導くことができる．

・従来の研究と本研究との関連性

判別分析については，Lachenbruch [4]，Lachenbruch and Mickey [5] がLOOを実際の誤判別率の推定や信頼区間の推定に使用している．Fukunaga and Kessell [2] はLOOが二次判別分析にも有効であることを述べている．Efron [1] はLOOを含むいくつかの手法に対し，実際の誤判別率の推定のバイアスを考察している．

判別分析におけるCVの理論研究は，重回帰分析に比べて多くない．そのため，特に応用の場面において，重回帰分析の研究結果をそのまま利用することがある．ところが，重回帰分析の評価基準は残差であり判別分析の評価基準とは異なる．この評価基準の違いにより，重回帰分析の研究結果を判別分析に読み替えることは問題と考える．たとえば，Hastie et al. [3, p.216] は，判別分析の場合についても 5, 10-fold CVがLOOに比べて優れていると述べている．しかし，著者の知る限り，明確に示した研究はない．実際，中西[6] が最終的に示した結論は，これを否定するものであった．

【一つ取って置き法（LOO）による実際の誤判別率の信頼区間】

２群の判別手法における誤判別率の推定を行うことを考える．

各観測を $z_j \ (j=1,\ldots,n)$とする．

LOOによる実際の誤判別率の推定値を，

$$err^{CV} = \frac{1}{n}\sum_{j=1}^{n} z_j \qquad (z_j = 1：\text{誤判別された場合}, z_j = 0：\text{正しい場合})$$

LOOによる実際の誤判別率の推定値の期待値を，

$$Err^{CV} = E[err^{CV}]$$

とおくと，LOOの推定がもたらす効果は次のような2つの要素で示すことができる（Hastie et al. [3, p.198]）．

$$E(Err - err^{CV})^2 = (Err - Err^{CV})^2 + E(Err^{CV} - err^{CV})^2$$

右辺の第１項を Biasの2乗，第２項をCVのVarianceと呼ぶ．LOOの推定においてはBiasの2乗＝0 とみなせる（詳細は省略）．あらためて．上の式は次のように近似式として表すことができる．

$$E(Err - err^{CV})^2 = E(Err^{CV} - err^{CV})^2$$

つまり，CVのVarianceは

$$V(\sum_{j=1}^{N} z_j)/N^2 = [\sum_{j=1}^{N} V(z_j) + \sum\sum_{k\neq l} \mathrm{cov}(z_k, z_l)]/N^2$$
$$= Err(1-Err)[1+(N-1)\rho]/N$$

と書き表すことができる（Lachenbruch [4]）．ここで，ρ は z_k と z_l の相関係数である．つまり，CVのVariance の信頼区間の近似として

$$\boxed{\Pr[(Err - err^{CV})^2 / \text{CVのVariance} \geq Z_{\alpha/2}^2] \leq 100\alpha/2}$$

が得られる．ここで，$Z_{\alpha/2}$ は標準正規分における上側確率 α/2点の値である．相関係数 ρ は判別ルール，データサイズ，変数の数などによって異なることが予想される．相関係数 ρ の評価についてはいくつかの数値例において評価した結果，次のようになった．

$$\boxed{E(Err - err^{CV})^2 < Err(1-Err)[1+0.5]/N}$$

ここまでの考察は Lachenbruch [4] が示したこととほぼ同じである．

【実際の誤判別率からみられる現実の問題】

LOOの実際の誤判別率の信頼区間が大きいことは，実際の判別ルール選択の問題となる．たとえば，表にあるルールA，B，Cからどれを選ぶべきかを LOOで判断しているとする．各ルールの実際の誤判別率の真値がそれぞれ0.2，0.3，0.4であるので，本来ならルールAを選びたいところであるが，LOOの信頼区間が大きいため，常に最適なルールが選ばれるわけではない（表の太文字を比較するとわかる）．これが判別分析において，CVが最適なルールを選ばない最大の理由である．

一つ取って置き法（LOO）における信頼区間の比較

ルール	Err	$(N-1)\ \rho$	N	$Var.$	67%信頼区間	95%信頼区間
A	0.2	0.3	40	0.072^2	(0.13, **0.27**)	(0.06, **0.34**)
B	0.3	0.3	40	0.083^2	(**0.22**, 0.38)	(**0.13**, 0.47)
C	0.4	0.3	40	0.088^2	(0.31, 049)	(**0.22**, 0.58)
A	0.2	0.5	40	0.077^2	(0.12, **0.28**)	(0.05, **0.35**)
B	0.3	0.5	40	0.089^2	(**0.21**, 0.39)	(**0.12**, 0.48)
C	0.4	0.5	40	0.095^2	(0.30, 0.50)	(**0.21**, 0.59)

中西[6] には，一般のK-fold CVについて考察を進め，LOOの性質の良さが示されている．

＜参考文献＞

[1] Efron, B. (1986) : “How Biased is the Apparent Error Rate of a Prediction Rule?”, JASA, 81, 461-470.

[2] Fukunaga, K. and Kessell, D. (1971) : “Estimation of classification error”, IEEE Trans. Computers, C20, 1521-1527.

[3] Hastie, T., Tibshirani, R. and Friedman, J. (2001) : The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Springer.

[4] Lachenbruch, P. A. (1967) : “An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis”, Biometorics, 23, 639-645.

[5] Lachenbruch, P. A. and Mickey, M. R. (1968) : “Estimation of error rates in discriminant analysis”, Technometorics, 10, 1-11.

[6] 判別分析におけるクロスバリデーションの信頼性(2008) : 成蹊大学経済学部論集，39(1)，85-102.