La Revue de Modulad

ESTIMATION DE LA QUALITE D'UNE REGLE DISCRIMINANTE

Gilles CELEUX
INRIA Rocquencourt

Jean-Christophe TURLOT
Université de Pau

Résumé :

On présente les différentes techniques de rééchantillonnage utilisées pour estimer le taux d'erreur d'une règle de décision en discrimination : échantillon test, bootstrap, jackknife, validation croisée. On passe en revue les possibilités offertes actuellement par les programmes de Modulad concernant l'utilisation de ces techniques et on indique les améliorations qui seront apportées dans la prochaine version de Modulad.

Mots-clés : discrimination, rééchantillonnage, test, bootstrap, validation croisée

1. Introduction

Sept programmes de la bibliothèque Modulad relèvent de la discrimination. Deux programmes, FUWIL et SELDISC, concernent exclusivement la sélection de variables pour la discrimination linéaire. Le programme ADM relève à la fois de la sélection de variables et de l'affectation d'un échantillon décrit par des variables continues et des variables qualitatives. Quatre programmes ont pour objet l'élaboration d'une règle de classement ; les programmes DIS2G et DISC construisent une règle de discrimination linéaire pour respectivement deux populations et plus de deux populations; les programmes SEGCLA et DNP relèvent de la discrimination par arbre, les variables explicatives (prédicteurs) étant respectivement qualitatives et quantitatives.

Il est bien sûr important de mesurer la qualité de ces règles de décision, à savoir le taux d'erreur de classement que l'on peut en attendre. Cet article est consacré à la présentation des techniques les plus utilisées pour fournir une estimation de ce taux d'erreur et il dresse un panorama de la situation actuelle dans la bibliothèque Modulad. Dans un souci de simplicité, nous n'évoquons pas le problème délicat de validation d'une règle de décision après sélection de prédicteurs. Ce sujet fait l'objet d'un autre article dans la présente revue (Turlot 1989). L'exposé fait référence aux notices d'emploi des programmes de la bibliothèque Modulad (Modulad 1987), ainsi qu'au document de l'Ecole Modulad sur l'analyse discriminante (Strasbourg 1989), en particulier aux présentations de Celeux : discrimination sur variables quantitatives (Modulad 1989a), de Daudin : discrimination logistique (Modulad 1989b), et de Turlot : validité de la règle de classement en discrimination (Modulad 1989c).

2. Estimation des taux d'erreur de classement

2.1. Rappels

On considère K populations $P_1, \ldots, P_K$, chacune d'entre elles étant décrite par un échantillon aléatoire d'observations. Les unités statistiques constituant le fichier de données appelé fichier d'**apprentissage** sont définies sans ambiguité par leur appartenance à l'une des populations P_k, $1 \leq k \leq K$ et par un vecteur x de p prédicteurs. On note y la variable indicatrice de la classe d'appartenance des observations. On rappelle que l'objet de l'analyse discriminante est la construction d'une règle de décision qui minimise le taux d'erreur de classement d'observations futures décrivant les populations potentielles de l'étude. Dans la suite, on note Err cette quantité et on note $Err(\ell/k)$ les probabilités conditionnelles d'erreur de classement d'un individu dans la population P_ℓ alors qu'il appartient à la population P_k.

Nous allons tout d'abord considérer le cas de la discrimination linéaire ; en effet, il s'agit de la technique de discrimination la plus utilisée lorsque les prédicteurs sont quantitatifs et elle donne dans la plupart des situations des résultats satisfaisants. On suppose que les vecteurs de prédicteurs suivent une loi normale non dégénérée de moyenne μ_k et de covariance Σ sur P_k $(1 \leq k \leq K)$. Si les coûts des différentes erreurs de classement sont égaux et si les probabilités a priori d'appartenance des observations à chacune des classes sont égales, la règle de classement est la suivante (Modulad 1989a) : Une observation décrite par le vecteur x de prédicteurs est affectée à la population P_k si et seulement si

$$(1) \qquad {}^t[x - \frac{1}{2}(\mu_k + \mu_\ell)]\, \Sigma^{-1}\, (\mu_k - \mu_\ell) \geq 0, \text{ pour tout } \ell \in \{1, \ldots, K\}$$

$^t u$ étant le vecteur transposé de u

En pratique, les paramètres (μ_k, $1 \leq k \leq K$ et Σ) du modèle sont inconnus et la règle de décision est estimée en remplaçant ces paramètres par leurs estimations du maximum de vraisemblance (g_k, $1 \leq k \leq K$ et S) dans l'expression (1); les g_k sont les centres de gravité des classes P_k et S est la matrice de variance intraclasse de la partition a priori $P=(P_1, ...,P_K)$. Ainsi, la règle de décision est conditionnelle au fichier d'apprentissage (y_1, x_1), ..., (y_n, x_n), noté (y, x) et qui est une réalisation d'un échantillon du vecteur aléatoire (Y,X) de loi F dont les paramètres sont inconnus. Le problème posé est alors celui de l'estimation des taux d'erreur de classement associés à cette règle de décision.

La probabilité $Err[\ell/k]$ d'affectation à la population $P\ell$ d'une observation future (ou potentielle) de la population P_k est appelée taux vrai d'erreur de classement. Pour expliciter l'expression de $Err[\ell/k]$, il est utile de remarquer que la règle de décision peut être caractérisée par K régions de l'espace $\mathbf{R}^p$ des prédicteurs, séparées deux à deux par des hyperplans (Modulad 1989a). On note R_k (y, x) la région d'affectation d'une observation quelconque (y_0, x_0) à la classe P_k, déterminée sur le fichier d'apprentissage (y, x). Pour ne pas alourdir les notations, on n'explicitera pas toujours le caractère conditionnel de la règle de classement. On a :

$$Err[\ell/k] = \text{Prob}\,[X \in R\ell / P_k], \tag{2}$$

autrement dit, il s'agit de la probabilité de la région $R\ell$ calculée selon la loi $N(\mu_k, \Sigma)$.

Ainsi, dans le cas de deux populations à discriminer, si les coûts des différentes erreurs de classement sont égaux et si les probabilités a priori d'appartenance des observations à chacune des classes sont égales, on obtient

$$Err[2/1] = \Phi\left\{\frac{-[\mu_1 - 1/2\ {}^t(g_1 + g_2)]\ S^{-1}\ (g_1 - g_2)}{{}^t(g_1 - g_2)\ S^{-1}\Sigma S^{-1}\ (g_1 - g_2)}\right\}$$

(3)

$$Err[1/2] = \Phi\left\{\frac{-[\mu_2 - 1/2\ {}^t(g_1 + g_2)]\ S^{-1}\ (g_2 - g_1)}{{}^t(g_1 - g_2)\ S^{-1}\Sigma S^{-1}\ (g_1 - g_2)}\right\}$$

où Φ représente la fonction de répartition d'une loi normale centrée et réduite. Il est facile de voir que $Err[2/1] = Err[1/2]$ (Anderson 1984). Il apparaît que ces quantités ne peuvent être calculées puisque les μ_k et Σ sont inconnus. On ne dispose que de résultats asymptotiques (Anderson 1984) fondés sur les hypothèses décrites plus haut et qui ont servi à l'élaboration de la règle de classement. Maintenant, si l'on remplace les paramètres μ_k et Σ par leurs estimations g_k et S sur le fichier d'apprentissage (y, x), on déduit de (3) l'estimation suivante des taux d'erreur de classement

$$err_D[1/2] = err_D[2/1] = \Phi\left(-\frac{D}{2}\right)$$

où

$$D^2 = {}^t(g_1 - g_2)\, S^{-1}\, (g_1 - g_2)$$

représente l'estimateur empirique de la distance de Mahalanobis Δ^2 entre les deux populations définie par

$$\Delta^2 = {}^t(\mu_1 - \mu_2)\, \Sigma^{-1}\, (\mu_1 - \mu_2)$$

Cette méthode de substitution conduit en général à une sous-estimation des taux d'erreur de classement qui peut être importante lorsque la taille du fichier d'apprentissage est modérée.

Mais, le point de vue adopté ici est différent : une règle de décision est généralement construite en faisant appel à des hypothèses plus ou moins adaptées (sous les hypothèses de la discrimination linéaire, les régions d'affectation $\{R_k,\ 1 \le k \le K\}$ sont séparées deux à deux par des hyperplans) ; de manière plus générale, on peut faire appel à d'autres hypothèses sur la loi de X et sur le lien entre la variable aléatoire Y décrivant l'appartenance à l'une des K classes et le vecteur aléatoire X, la règle de décision prenant alors une autre forme, comme en discrimination logistique (Modulad 1989b). Mais ce qui importe est l'évaluation de la qualité de la règle de classement obtenue ; pour celà, on souhaite s'affranchir des hypothèses les plus fortes ayant servies à l'élaboration de la règle de classement.

Deux types de méthodes non paramétriques d'estimation des taux d'erreurs de classement sont présentées : l'utilisation d'un fichier test (ou fichier d'épreuve) et les techniques de rééchantillonnage par validation croisée et bootstrap (Modulad 1989c).

2.2 L'échantillon test

Un estimateur naturel des taux $Err[\ell/k]$ est obtenu en appliquant la règle de classement au fichier d'apprentissage ; soit $err[\ell/k]$ la fréquence des observations de la classe P_k affectées à la classe $P\ell$. Cette quantité est appelée taux **apparent** d'erreur. Comme le fichier est utilisé à la fois pour l'élaboration de la règle de classement et pour son évaluation, un tel estimateur est généralement trop **optimiste** : les vrais taux d'erreurs de classement sont nettement sous-estimés lorque la taille de l'échantillon est modérée. On note $op[\ell/k]$ ces optimismes et op le biais (l'optimisme) inconditionnel de l'estimation err de Err :

$$Err[\ell/k] = err[\ell/k] + op[\ell/k] \qquad 1 \le \ell, k \le K\ ;\ \ell \ne k$$

(4)

$$Err = err + op$$

les quantités $op(\ell/k)$ et op sont évidemment inconnues.

Cependant, si Err est estimée sur un fichier $(\mathbf{y}_0, \mathbf{x}_0)$ indépendant du fichier d'apprentissage, op est d'espérance nulle ; en d'autres termes, les taux apparents d'erreur calculés sur le fichier $(\mathbf{y}_0, \mathbf{x}_0)$ appelé fichier **test** (ou fichier d'épreuve) sont sans biais.

En pratique, ce fichier est construit par tirage aléatoire d'un sous-ensemble d'unités statistiques du fichier (de l'ordre de 10 à 50 % des unités statistiques du fichier d'origine suivant sa taille), les données restantes constituant le fichier d'apprentissage. Bien entendu, une telle procédure n'est possible que si le nombre d'observations est suffisamment important dans chacune des populations. L'avantage de cette technique est l'absence de biais des estimateurs des taux d'erreur de classement, leur précision dépendant de la taille du fichier test. En contrepartie, la règle de décision est estimée de

manière moins précise en ce sens que l'information n'y est pas utilisée de manière exhaustive. Aussi, il convient, après avoir évalué la qualité de la règle de décision à l'aide d'un échantillon test, de proposer la règle de décision construite sur la totalité du fichier disponible.

2.3. Les techniques de rééchantillonnage

On ne dispose pas toujours d'un fichier de taille suffisante pour en tirer un échantillon d'apprentissage et un échantillon test de tailles raisonnables. On a alors recours aux techniques de rééchantillonage appliquées au fichier $(\mathbf{y}, \mathbf{x})$ pour évaluer op et les op(ℓ/k). Les techniques les plus connues sont la validation croisée, le jackknife et le bootstrap.

Le principe de la **validation croisée** consiste à retirer successivement chacune des observations (y_i, x_i) du fichier d'apprentissage, à estimer la règle de classement sur le nouveau fichier $(\mathbf{y}, \mathbf{x})_{(i)}$ de taille n - 1 (n étant le nombre d'observations), puis à affecter l'observation absente (y_i, x_i) selon cette règle. On en déduit des estimations $err_{CV}[\ell/k]$ des taux d'erreur conditionnels qui sont définies par la fréquence des observations de la classe P_k affectées à la classe P_ℓ selon cette procédure. Enfin, le taux moyen d'erreur est estimé par

$$(5) \qquad err_{CV} = \sum_{k=1}^{K} \sum_{\ell \neq k} p_k \, err_{CV}(\ell/k)$$

où p_k est le poids de la population P_k.

En d'autres termes, le biais op de l'estimateur err du taux vrai d'erreur de classement Err associé à la fonction de décision construite sur le fichier d'apprentissage $(\mathbf{y}, \mathbf{x})$ est estimé par la quantité $b_{CV} = err_{CV} - err$.

Une autre technique de rééchantillonnage, appelée **bootstrap** peut être appliquée pour l'estimation du biais. Elle consiste à affecter une probabilité 1/n à chacune des observations du fichier d'apprentissage, puis à tirer aléatoirement - avec remise - un échantillon de taille n, appelé fichier bootstrap. Un fichier bootsrap est ainsi obtenu en remplaçant la loi F inconnue du couple (Y, X) dont est issu le fichier d'apprentissage $(\mathbf{y}, \mathbf{x})$ par sa distribution **empirique** F'. On peut ainsi construire un grand nombre de fichiers bootstrap $(\mathbf{y}, \mathbf{x})_\alpha$ $1 \leq \alpha \leq A$, par tirages successifs de A échantillons de taille n.

Sur chacun de ces fichiers, on élabore la règle r_α de classement (on estime les paramètres du modèle de discrimination choisi), puis on calcule les $err_\alpha[\ell/k]$, à savoir les taux d'erreur apparents calculés sur le fichier $(\mathbf{y}, \mathbf{x})_\alpha$ et les $Err_\alpha[\ell/k]$, c'est-à-dire les estimations bootstrap des taux d'erreur réels de la règle r_α calculé sur le fichier initial $(\mathbf{y}, \mathbf{x})$. Finalement, l'estimation bootstrap du biais est défini par la moyenne des écarts entre les quantités $Err_\alpha[\ell/k]$ et $err_\alpha[\ell/k]$ sur l'ensemble des A échantillons bootstrap, de sorte que les vrais taux d'erreur de classement associés à la règle de décision fondée sur le fichier d'apprentissage sont estimés par les quantités :

$$(6) \qquad err_b[\ell/k] = err[\ell/k] + 1/A \sum_{\alpha} \{Err_\alpha[\ell/k] - err_\alpha[\ell/k]\}, \ \ell \neq k.$$

Le taux inconditionnel d'erreur est estimé par

$$err_b = err + 1/A \sum_{\alpha} \sum_{\ell} \sum_{k \neq \ell} p_k \{Err_\alpha[\ell/k] - err_\alpha[\ell/k]\}.$$

D'autres techniques de rééchantillonnage ont été proposés pour l'estimation du biais des taux d'erreur de classement, comme le jackknife qui conduit à une approximation de l'estimateur bootstrap (Efron 1983). D'un point de vue plus formel, il semble que l'usage des différentes techniques de rééchantillonnage pour l'estimation des taux d'erreur d'affectation repose sur une même idée : la substitution de la loi théorique F, inconnue, du couple (Y, X) dont est issu le fichier (**y**, **x**) par une loi F* qui lui est "proche", et permettant d'estimer le biais b.

De manière plus précise, les quantités err(**y**, **x**) et op(**y**, **x**), liées entre elles par la relation (4) dépendent de la règle de classement estimée sur le fichier d'apprentissage, et sont donc conditionnelle à (**y**, **x**).

Soit E_F [op(Y, X)] l'espérance de op par rapport à la loi F du couple (Y, X), on a

$$b = E_F [op(Y, X)] = E_F [Err(Y, X) - err (Y, X)] \quad (7)$$

Si la quantité b était connue, un estimateur naturel de Err(y, x) serait donné par la quantité err (**y**, **x**) + b. En pratique, b doit être estimé à partir du fichier d'apprentissage ; en subsistant à la loi F sa loi empirique F', on peut approcher la quantité $E_{F'}$ [op(Y', X')] par la technique bootstrap. On en déduit l'estimation bootstrap de Err(**y**, **x**) en corrigeant le taux apparent d'erreur err(**y**, **x**) par l'estimation de b' = $E_{F'}$ [op(Y', X')]. Sous cet éclairage, l'estimation bootstrap paraît naturelle. Comme l'estimateur de b obtenu par la technique Jackknife est une approximation quadratique de l'estimateur bootstrap (Efron 1983) et que l'estimateur par validation croisée en est très proche dans sa forme (Efron 1982), tout celà indique qu'une même idée est à la base de l'application des techniques de rééchantillonnage : la substitution de F, inconnue par une distribution F*, de manière à pouvoir estimer b(F*) = E_{F*} (op).

Cette présentation montre clairement que la quantité op(**y**, **x**) est remplacée par une estimation de b = E_F [op (Y, X)] dans l'expression (3), c'est-à-dire dans l'estimation de Err(**y**, **x**). Celà n'est pas sans conséquence, puisque, si la loi de op (Y, X) est très dispersée, l'écart entre Err[**y**, **x**] et son estimation peut être important.

Efron (1983) a montré sur des simulations que les techniques bootstrap et de validation croisée conduisaient à des estimations satisfaisantes du taux vrai d'erreur de classement. De manière plus précise, ces résultats montrent que l'estimation de b par validation croisée (b_{CV} = err_{CV} - err) présente un biais plus petit que l'estimateur bootstrap b' ; ils montrent aussi que la variabilité du premier est nettement plus importante que celle du second, de sorte que l'estimateur bootstrap apparaît plus précis. Il faut cependant noter que ces simulations reposent sur le modèle normal dans le cas de deux populations P_1 et P_2 ; un tel modèle ne recouvre évidemment pas la diversité des données rencontrées en pratique, ce qui limite la portée de ces résultats.

Un défaut évident de ces techniques est leur consommation très importante de temps calcul. De ce point de vue, la validation croisée est hautement préférable au bootstrap. Le fait qu'à chaque étape de la validation croisée le fichier d'apprentissage n'est amputé que d'une observation permet la plupart du temps de modifier la règle de décision sans avoir à recommencer tous les calculs. Ainsi, on expose dans l'annexe comment effectuer rapidement la mise à jour de la règle de discrimination linéaire par validation croisée. Ce genre de simplification est tout à fait impossible avec le bootstrap car à chaque étape l'échantillon d'apprentissage est profondément modifié.

Enfin, on doit noter que l'implantation de la validation croisée pour définir des règles optimales d'élagage des arbres de décision (Breiman et al. 1984) constitue une

avancée très remarquable pour la discrimination par arbre. Par contre, dans ce cadre, le bootstrap s'avère impraticable.

3. La situation actuelle dans la bibliothèque Modulad

Le programmes SEGCLA ne possède pas de procédure de validation des règles de décision.

Le programme ADM fournit des estimations des Err[ℓ/k] sur la base des hypothèses paramétriques utilisées pour construire les règles de décision, mais ne possède pas de procédure non paramétrique d'évaluation de ces règles.

Le programme DIS2G de discrimination linéaire pour deux populations propose une estimation des taux d'erreur de classement par échantillon test et par bootstrap.

Le programme DISC de discrimination linéaire pour un nombre quelconque de populations propose une estimation des taux d'erreur de classement par échantillon test et par validation croisée. On doit signaler que l'implatation de la procédure de validation croisée est toute récente et n'est pas disponible dans la version 2.2 de Modulad, mais le sera dans la version 2.3.

Les programmes de sélection FUWIL et SELDIS ne proposent pas de procédure de validation de la sélection.

Pour les programmes de discrimination linéaire DIS2G et DISC, les estimations possibles des taux d'erreur de classement associés à la fonction linéaire discriminante élaborée sur le fichier d'apprentissage sont éditées ainsi :

- les taux apparents de classement : err et err(ℓ/k) ; $1 \leq \ell, k \leq K$;
- les estimations de Err et de Err(ℓ/k), $1 \leq \ell, k \leq K$ fondées sur un échantillon test (en option) ;
- les estimations bootstrap (en option) de Err et de Err(ℓ/k) ; $1 \leq \ell, k \leq K$, dans le cas de deux populations [programme DIS2G] ;
- les estimations par validation croisée (en option) de Err et de Err(ℓ/k) ; $1 \leq \ell, k \leq K$, dans le cas d'un nombre quelconque de populations [programme DISC].

Les résultats se présentent sous la forme de tableaux [K x K] : en lignes figurent les classes vraies d'appartenance ; en colonnes les classes d'affectation. Sur la ligne k ($1 \leq k \leq K$) sont données les fréquences d'affectation aux différentes classes des unités statistique de la population P_k, de sorte que les fréquences de reclassement corrects estimés figurent sur la diagonale des tableaux.

En pratique, la qualité de la règle de décision est estimée soit par le taux moyen d'erreur de classement sur l'échantillon test, soit par son estimation par rééchantillonnage, mais en aucun cas par le taux apparent d'erreur. Cependant, il est toujours utile de comparer le tableau des taux apparents d'erreur à celui obtenu par la technique de validation choisie, car l'écart entre les résultats montre l'importance des fluctuations d'échantillonnage dans l'estimation des paramètres de la fonction linéaire discriminante.

4. Conclusion

Les modèles paramétriques de discrimination permettent de définir des régions de décision dans l'espace des prédicteurs, dont la forme est déterminée par les hypothèses faites sur la loi F du vecteur aléatoire (Y, X) dont est issu le fichier des données. On peut s'affranchir de ces hypothèses plus ou moins adaptées à la nature des données pour estimer la qualité de la règle de classement élaborée sur le fichier d'apprentissage en ayant recours à des techniques non paramétriques.

Si de nombreuses données ont été observées, le plus sûr est de construire aléatoirement un fichier test, puis d'estimer les taux d'erreur de classement sur ce fichier, ces estimateurs étant sans biais. Lorsque la taille de l'échantillon est modérée, on doit faire appel à des techniques de rééchantillonnage comme le bootstrap ou la validation croisée ; elles conduisent à des estimations satisfaisantes des taux d'erreur lorsque le choix des prédicteurs est effectué a priori.

Toutefois, la construction des règles de décision est en général précédée d'une procédure de sélection d'un sous-ensemble pertinent de variables prédictives (Modulad 1989a et c). Cette sélection introduit un biais dans l'évaluation du taux d'erreur qu'il convient de prendre en compte. Turlot dans l'article qu'il consacre à cette question dans le présent numéro de La Revue de Modulad donne des éléments de réponse à cette question (Turlot 1989).

Les programmes de Modulad concernant la discrimination linéaire sont assez complets pour évaluer les règles de décision si l'on tient compte de la nouvelle version de DISC qui incorpore une procédure de validation croisée. Par contre, on peut penser que les procédures de sélection (FUWIL, SELDIS et ADM) gagneraient à proposer des techniques de validation. De même, les programmes de discrimination par arbre (SEGCLA, DNP) sont peu satisfaisant sous cet aspect. Mais un programme, inspiré du programme CART de Breiman et al. (1989) prenant simultanément en compte des prédicteurs quantitatifs et qualitatifs et proposant une procédure optimale d'élagage par validation croisée est annoncé ....

Annexe

Mise à jour de la règle de discrimination linéaire par validation croisée : éléments permettant cette mise à jour rapidement.

La règle de discrimination linéaire nécessite le cacul des centres de gravité g_k des populations P_k, le calcul de la matrice variance intraclasse S de la partition $P = (P_1, \ldots, P_K)$ et le calcul de la matrice inverse S^{-1}. Nous allons voir comment actualiser ces quantités, lorsque l'on ampute le fichier d'apprentissage (y,x) de l'observation (y_i,x_i) $1 \leq i \leq n$, n étant la taille du fichier. On notera n_k le cardinal de la classe P_k.

1 - *Actulisation des centres de gravités*

Notons (i) la classe d'appartenance de l'observation i. Seul le centre de gravité $g_{(i)}$ est modifié. Des calculs simples montrent qu'il devient

$$g'_{(i)} = \frac{n_{(i)}g_{(i)} - x_i}{n_{(i)} - 1}$$

2 - *Actualisation de la matrice variance intraclasse*

La matrice variance intraclasse peut s'écrire $S = 1/n \sum_{k=1}^{K} n_k S_k$ où les S_k représentent la matrice variance de la classe P_k. Dans cette formule seule la matrice $S_{(i)}$ est modifiée. On peut montrer qu'elle devient

$$S'_{(i)} = \frac{n_{(i)}S_{(i)}}{n_{(i)} - 1} - n_{(i)} \; (g_{(i)} - g'_{(i)}) \; {}^t(g_{(i)} - g'_{(i)})$$

On en déduit que la matrice S devient

$$S' = \frac{nS - (n_{(i)} - 1)n_{(i)}(g_{(i)} - g'_{(i)}) \; {}^t(g_{(i)} - g'_{(i)})}{n - 1}$$

3 - *Inversion de S*

D'après la dernière équation, on voit que S' est de la forme $A + u{}^tu$, A^{-1} étant connu et u étant un vecteur. On en déduit (cf. Hager 1989, par exemple) que

$$S'^{-1} = A^{-1} - \frac{A^{-1}u{}^tuA^{-1}}{1 + {}^tuA^{-1}u}$$

Finalement, il s'ensuit que l'on a tous les éléments pour modifier rapidement la règle de discrimination linéaire après avoir supprimé un élément du fichier d'apprentissage. En particulier, il n'est pas nécessaire d'inverser à chaque fois la matrice variance intraclasse S.

Références

Anderson T. W. (1984) *An introduction to Multivariate Statistical Analysis* - Wiley.

Breiman L., Friedman J.H., Ohlsen R.A., Stone C.J (1984) *Classification and regression trees* - Wadsworth.

Daudin J.J., Soukal M. (1988) Analyse discriminante sur variables continues et qualitatives, notice scientifique du logiciel ADM - *La Revue de Modulad* **1**. 1-12.

Efron B. (1982) *The Jackknife, the Bootstrap and other resampling plans* - SIAM Monograph **38**.

Efron B. (1983) Estimating the error rate in a predictive rule : improvement on cross validation - *JASA* **78**. 316-331.

Hager W. W. (1989) Updating the inverse of a matrix - *SIAM Review* **31**. 221-239.

Modulad (1987) *Bibliothèque FORTRAN pour l'analyse des données* - INRIA, (Supplément 1989).

Modulad (1989) *Méthodes de Discrimination* - Support de cours de l'école Modulad (Strasbourg 1989), INRIA.

a- Celeux G. Discrimination sur variables quantitatives.
b- Daudin J.J. Discrimination logistique.
c- Turlot J.C. Validité de la règle de classement en discrimination.

Turlot (1989) Sélection de prédicteurs et estimations des taux d'erreur en discrimination linéaire - *La Revue de Modulad* **4**. 47-60.