L'ANALYSE DISCRIMINANTE SUR VARIABLES QUALITATIVES :

PROGRAMMES DE QUATRE METHODES

Abdallah Mkhadri

Abdelaziz Nasroallah

Département de Mathématiques
Faculté des Sciences Semlalia
B.P. S 2390
Marrakech, Maroc
mkadri@ucam.ac.ma
nasroallah@ucam.ac.ma

Résumé : *Nous présentons quatre modèles de discrimination sur variables qualitatives : la discrimination prédictive fondée sur le modèle d'indépendance conditionnelle, deux modèles graphiques décomposables et le modèle logistique linéaire. Toutes ces méthodes visent à réduire la complexité du modèle multinomial complet. Nous introduisons ensuite, les programmes des quatre méthodes.*

Mots-clefs : *analyse discriminante, discrimination prédictive, variables qualitatives, indépendance conditionnelle, modèles graphiques décomposables, modèle logistique.*

Les programmes (exécutables et données) sont sur le WEB à l'adresse de la revue. Pour toute information concernant ces programmes veuillez contacter nasroallah@ucam.ac.ma.

1. Introduction

On dispose d'un échantillon de n individus décrits par p variables qualitatives X^1, ..., X^p dont l'appartenance à l'un des k groupes a priori $G_1,...,G_k$ est connu. L'ensemble **E** désignant l'espace des observations (par exemple si les variables sont toutes binaires, alors $\mathbf{E}=\{0,1\}^p$), nous notons par $\mathbf{Z}^n = \{(x_1,y_1),...,(x_n,y_n)\}$ l'échantillon d'apprentissage et par $G = (G_1,...,G_k)$ la partition définie sur $\mathbf{Z}^n$, où :

- x_i élément de **E** est le vecteur de description de la $i^{ème}$ observation par les p variables
- y_i appartenant à $\{1,...,k\}$ indique le groupe auquel appartient cette observation.

Le problème est alors de construire, sur cet échantillon, des règles de décision qui produiront un minimum d'erreurs lorsqu'elles seront appliquées dans le futur. Plus précisément, on suppose que la partition G est munie d'une loi de probabilité $\Pi = (\Pi_1, ..., \Pi_k)$ appartenant à S_k, où S_k est le simplexe unité de 3^k. On suppose de plus que les points de **E** provenant d'un même groupe G_h sont les réalisations d'une variable aléatoire de distribution P_h par rapport à la mesure de comptage sur **E**, nous posons $\mathbf{P} = (P_1,...,P_k)$. Dans ce cadre, toute règle de décision θ_n : $\mathbf{E}x\mathbf{S}_k x\mathbf{Z}^n \rightarrow \{1,...,k\}$ est une fonction mesurable dépendante de la valeur de l'observation **x**, du vecteur de probabilité a priori Π et de $\mathbf{Z}^n$. On considère le problème selon le point de vue de la théorie de décision et dans ce cas l'espace d'action et l'espace des paramètres est $\{1,...,k\}$. Une fonction de perte arbitraire L est une fonction à valeur réelle sur $\{1,...,k\}x\{1,...,k\}$; $L(i,j)$ représente le coût de mauvais classement d'un individu (i.e. une observation) de G_j dans G_i pour i, j = 1, ..., k. Par ailleurs, la probabilité a posteriori du groupe i est donnée par :

$$P^*[\theta = i \mid X = x] = \frac{\pi_i P_i(x)}{\sum_j \pi_j P_j(x)}$$

Le risque de Bayes d'une fonction de décision δ est définie comme l'espérance du coût de cette fonction de décision par rapport à la probabilité a posteriori P^*, i.e.

$$R_{P^*}(\delta, x) = \sum_i L(\delta, i) P^*[\theta = i \mid X = x]$$

La règle de décision bayésienne, $\hat{\theta}_B(x)$, est celle qui minimise le risque $R_{P^*}(\delta, x)$, i.e. :

$$R_{P^*}(\hat{\theta}_B, x) = \min_{\delta(x) \in \{1,\ldots,k\}} R_{P^*}(\delta, x)$$

Dans le cas particulier de la fonction de perte 0-1 qui est la plus utilisée (i.e. L(i,i) = 0 et L(i,j) = 1, $i \neq j$), la règle de Bayes, qui réalise ce minimum, consiste à :

$$\text{affecter } \mathbf{x} \text{ à } G_h \Leftrightarrow h = \arg\max_j \pi_j P_j(x) \qquad (1)$$

En pratique, les quantités en jeu dans le modèle bayésien ne sont pas connues et doivent être estimées. L'estimation des probabilités a priori ne relèvent pas de l'analyse statistique de l'échantillon d'apprentissage $\mathbf{Z}^n$, sauf si $\mathbf{Z}^n$ est issu d'un mélange de lois, auquel cas les probabilités a priori seront estimées par les proportions. Dans l'autre cas, leur choix peut être délicat et bien souvent on est amené à se placer sous l'hypothèse d'égalité des probabilités a priori. Finalement, l'estimation statistique d'une règle de décision revient à estimer les probabilités $P_j(\mathbf{x})$, $j = 1,\ldots,k$. Lorsque les variables sont qualitatives (i.e. à valeurs discrètes), il est naturel d'utiliser des distributions de probabilités discrètes et bien sûr de manière privilégiée la distribution multinomiale. Une manière de réduire la complexité de ce modèle est de considérer un modèle parcimonieux et robuste comme le modèle d'indépendance conditionnelle (MIC). MIC consiste à supposer que les p variables sont indépendantes dans chaque groupe. Ainsi l'estimation de la probabilité conditionnelle $P_h(\mathbf{x})$ s'écrit :

$$\hat{P}_h(x) = \prod_{j=1}^{p} \frac{\#\left\{x_i \in G_h \mid x_i^j = x^j\right\}}{n_h} \qquad (2)$$

désignant le cardinal d'un ensemble, x^j la j^{eme} composante du vecteur $\mathbf{x}$ et $n_h = \#\{x_i \in G_h\}$.

MIC fournit en général de bons résultats et il est considéré comme une méthode de référence dans ce cadre (cf. [CEL94] Ch.2). Bochi, Celeux et Mkhadri ([BOC93]) ont fourni à la bibliothèque Modulad le programme DISIND correspondant à MIC.

L'objet de cette notice est de présenter quatre nouveaux programmes associés à quatre nouvelles méthodes de discrimination discrète qui visent aussi à réduire le problème d'identification par la diminution du nombre de paramètres à estimer.

2. Discrimination prédictive discrète

L'approche prédictive de discrimination consiste à remplacer dans (1) les distributions de probabilité conditionnelle par les distributions de probabilité prédictive (cf. [GEI66], [GEI93]) qui s'écrivent sous la forme :

$$p_h(x \mid X^n) = \int_\Theta P_h(x \mid \theta_h)\psi(\theta_h \mid X^n)d\theta_h \qquad (3)$$

où $\psi(\theta_h \mid \mathbf{X}^n)$ est une certaine densité a posteriori et $X^n = \{x_i\}_{i=1}^n$. L'utilisation de la règle (1) par la discrimination prédictive a été justifiée récemment par Johnson et Mouhab ([JOH96]). Par ailleurs, cette méthode a été plutôt considérée uniquement dans le cas où les variables sont quantitatives : car en choisissant un modèle gaussien pour chaque groupe et une loi non informative de Jeffery pour les paramètres, on obtient une forme explicite pour la densité prédictive qui est une densité de Student (cf. [MKH97]). Maintenant, supposons que $\mathbf{E} = \{0,1\}^p$ et que chaque groupe G_h est représenté par un vecteur binaire $\mathbf{a}_h$. De même, nous notons par $n_{jh} = \sum_{x_i \in G_h} \left| x_i^j - a_h^j \right|$ le nombre d'observations du groupe G_h qui diffère de $\mathbf{a}_h$ pour la j^{eme} composante et alors $\theta_{jh} = n_{jh}/n_h$ représente la fréquence relative ($0 \le \theta_{jh} \le 1$). Gyllenberg et Koski ([GYL97]) ont montré que si $P_h(\mathbf{x} \mid \mathbf{a}_h, \theta_{jh})$ est une distribution de Bernoulli multivariée et si, θ_h est un vecteur de p variables aléatoires indépendantes suivant une loi Beta, alors la densité prédictive s'écrit :

$$p_h(x \mid X^n, a_h) = \prod_{j=1}^{p} \left(\frac{n_{jh}+1}{n_h+2} \right)^{w(j,h)} \left(1 - \frac{n_{jh}+1}{n_h+2} \right)^{1-w(j,h)} \qquad (4)$$

où $w(j,h) = | x^j - a_h^j |$. Nous avons remarqué que la preuve de ce résultat reste inchangée, dans le

cas de variables qualitatives, si on remplace la distance L_1 par la distance d définie par $d(x^q, y^q) = 1$ si $x^q \neq y^q$ et 0 sinon, où x^q (resp. y^q) est la modalité de la variable q pour l'observation $\mathbf{x}$ (resp. $\mathbf{y}$). Dans ce cas, la formule (4) devient :

$$p_h^*(x \mid X^n, a_h) = \prod_{j=1}^{p} \left(\frac{n_{jh}^*+1}{n_h+2} \right)^{d^*(j,h)} \left(1 - \frac{n_{jh}^*+1}{n_h+2} \right)^{1-d^*(j,h)} \qquad (5)$$

où $n^*_{jh} = \sum_{x_i \in G_h} d(x_i^j, a_h^j)$ et $d^*(j,h) = d(x^j, a_h^j)$. Comme pour la classification de données

binaires ou qualitatives ([GOV90], [NAD93]), le représentant de la classe a est obtenu soit par la règle majoritaire ou soit par maximisation de l'entropie ou soit par maximisation de la vraisemblance prédictive (cf. [GYL97]).

L'avantage de (5) est qu'elle n'est jamais nulle pour toute cellule, alors que l'expression (2) du MIC nécessite une correction de chaque terme du produit pour éviter ce problème (cf. [CEL94] ch.2). Par ailleurs, nous avons effectué une petite comparaison numérique des modèles (2) et (5) sur trois échantillons de données. Elle montre que les valeurs de (2) et (5) pour toute observation sont en général différentes, et que les décisions fournies par les deux modèles sont relativement identiques. Mais, il est possible que dans d'autres situations les deux modèles peuvent fournir des résultats différents.

Nous désignons par DIPIND le programme correspondant à la discrimination prédictive fondée sur le modèle d'indépendance conditionnelle (5).

3. Modèles graphiques décomposables

L'hypothèse d'indépendance conditionnelle, comme MIC ou DIPIND, est simple mais peut paraître non réaliste dans certaines applications. Comme alternative Chow et Liu ([CHO68]) proposent les modèles graphiques décomposables (ou réseaux probabilistes) qui fournissent pour chaque groupe à discriminer une estimation de sa densité de probabilité en tenant compte de certaines relations de dépendance conditionnelle entre variables explicatives. Plus précisément, chaque groupe G_h est associée à un réseau probabiliste défini par un couple (A, P_h) où A est un arbre fini associé à une distribution de probabilité P_h. Les sommets de l'arbre représentent les variables aléatoires $X^1,\ldots,X^p$; deux sommets sont reliés entre eux par une arête dont la longueur est une mesure de liaison. Ils proposent comme approximation de la distribution P inconnue le "champ markovien" P_h^I qui s'écrit :

$$P_h^I(x) = \prod_{j=1}^{p} P_h\,[x^j \mid x^{i(j)}] \qquad (6)$$

où $X^{i(j)}$ est la variable qui désigne le parent de X^j. La racine X^1 de l'arbre peut être arbitrairement choisie et caractérisée par la probabilité $P(x^1)$. Ce modèle graphique (6) est

estimé par minimisation de la distance de Kullback-Leibler entre la vraie distribution P et son approximation P^I :

$$d_{KL}(P,P^I) = \sum_h \sum_{x,z} P_h(x,z) \ln\left(\frac{P_h(x,z)}{P_h^T(x)P_h^T(z)}\right)$$

Les quantités $P_h(\mathbf{x},\mathbf{z})$ sont remplacées par les fréquences observées et l'arbre de dépendance est obtenu en utilisant un algorithme de Kruskal détaillé dans [CHO68], et [CEL94] ch. 4). Alors les probabilités correspondantes $P_h^I(x_i)$ sont estimées par les estimateurs de maximum de vraisemblance. *Nous désignons par DISARC le programme associée à la discrimination fondée sur les arbres de dépendance conditionnelle.*

Par ailleurs, dans une perspective parcimonieuse, Wong et Wang ([WON77]) suggèrent de considérer la forme (6) mais avec un seul arbre pour tous les groupes qui maximise

$$\sum_{i=1}^{p}\left[\sum_{h=1}^{k}\pi_h I_h(x^i,x^{i(j)}) - I(x^i,x^{i(j)})\right]$$

où $\mathbf{I}_h(x^i,x^{i(j)})$ (resp. $\mathbf{I}(x^i,x^{i(j)})$) est la mesure d'information mutuelle conditionnelle au groupe h (resp. globale) (cf. [MKH94]). *On désigne par DISARG le programme associé à la discrimination fondée sur un seul arbre de dépendance globale.*

Wong et Poon ([WON89]) affirment, d'après une petite étude par simulation, que si DISARG est attractif au niveau de calculs, il peut être moins efficace que DISARC. Mais, cette affirmation peut être critiquée spécialement dans le cas des échantillons de petites tailles. En effet, il est bien connu dans ce cadre que CIM, qui est un cas particulier de DISARG, fournit souvent de bonnes performances. Par ailleurs, une récente étude comparative menée par Nasroallah ([NAS98]) montre que DISARG fournit des résultats relativement meilleurs que DISARC; ce qui confirme encore cette critique. Finalement, nous pensons que ces deux modèles sont en fait très utiles lorsque la dimension est très grande.

4. Discrimination logistique

Le modèle logistique est un modèle semi-paramétrique qui est notamment utilisé dans le domaine biomédical et plus particulièrement dans les enquêtes épidémiologiques pour étudier l'association entre une maladie et un facteur de risque ([HOS89]). Les références essentielles

de son application en discrimination sont [DAY67] et surtout [AND72] et [AND75] (voir [CEL94] Ch.3 pour d'autres références). Généralement, le modèle logistique vise à décrire la liaison entre une variable qualitative Y à k modalités (variable de groupe) et l'ensemble de variables explicatives $(X^1,\ldots,X^p)$. Dans le cas de deux groupes à discriminer, supposons que $\mathbf{x}$ est un vecteur d'observation, alors le logarithme du rapport de vraisemblance est modélisé par

$$\ln\left[\frac{P_1(x)}{P_2(x)}\right] = \beta^t\mathbf{x}^* \qquad (7)$$

où $\mathbf{x}^* = (1,x^1,\ldots,x^p)^t$ et $\beta^t = (\beta_0^0, \beta_1,\ldots, \beta_p)$ est un vecteur de paramètres inconnus. Comme MIC la discrimination logistique linéaire est parcimonieuse : le nombre de paramètres à estimer est égal à p + 1. Maintenant, si nous posons :

$$\nu_j(\mathbf{x}) = P(Y = j \mid \mathbf{x}) \quad j = 1, 2, \quad \text{avec} \quad \nu_1(\mathbf{x}) + \nu_2(\mathbf{x}) = 1,$$

nous en déduisons alors, en calculant le rapport de $\nu_1(\mathbf{x})/\nu_2(\mathbf{x})$ par la formule de Bayes et en prenant le logarithme, que :

$$\ln\left[\frac{\nu_1(x)}{1-\nu_1(x)}\right] = \ln(\pi_1/\pi_2) + \beta^t\mathbf{x}^*.$$

Ce qui est équivalent à :

$$\nu_1(\mathbf{x}) = \frac{\exp(\beta_0 + \beta^{*t}\, x)}{1+\exp(\beta_0 + \beta^{*t}\, x)}$$

où $(\beta_0 = \ln(\pi_1/\pi_2) + \beta_0^0$ et $\beta^{*t} = (\beta_1,\ldots, \beta_p)$, donc :

$$\text{Logit}\left[\frac{\upsilon_1(x)}{1-\upsilon_1(x)}\right] = \beta_0+\beta^{*t}\mathbf{x} \qquad (8)$$

L'observation $\mathbf{x}$ est alors affectée au groupe 1 si $\text{Logit}[\nu_1(\mathbf{x})/(1 - \nu_1(\mathbf{x}))] > 0$ et au groupe 2 sinon.

Maintenant, pour définir le modèle logistique dans le cas de k groupes, nous notons par $\beta_s = \beta_{s1},\ldots, \beta_{sp})$, $s = 1,\ldots, k - 1$, le vecteur de paramètres. En considérant, par exemple, le groupe k comme le groupe de référence, le modèle s'écrit alors :

$$\text{Logit}[\nu_s(\mathbf{x})/\nu_k(\mathbf{x})] = \beta_{s0}+\beta_s^{*t}\mathbf{x}\ \} \qquad (9)$$

ce qui est équivalent à :

$$\nu_s(x) = \frac{\exp(\beta_{s0}+\beta_s^{*t}x)}{1+\sum_h \exp(\beta_{h0}+\beta_h^{*t}x)} \qquad (10)$$

où ($\beta_0 = \ln(\pi_s/\pi_k) + \beta_{s0}^0$, s = 1,..., k- 1. Les paramètres du modèle sont estimés par la méthode du maximum de vraisemblance en utilisant un algorithme de Newton-Raphson. Pour plus de détails sur le problème d'estimation voir par exemple [CEL94] Ch.3. *Nous désignons par DISLOG le programme correspondant à la discrimination logistique linéaire.*

5. Description des programmes

Les programmes DIPIND, DISARC, DISARG et DISLOG, sont écrits en langage Pascal, réalisent la discrimination sur variables qualitatives, selon les modèles respectifs décrits dans les sections précédentes, en privilégiant le point de vue décisionnel. Nous avons essayé de fournir la même interface que le programme DISIND disponible dans la bibliothèque de Modulad. En entrée, les variables explicatives sont qualitatives avec un nombre de modalités fixé selon la nature du programme, et le cas échéant ce nombre peut être modifié dans le programme source. Le fichier de données doit être en code ASCII où il est indispensable de laisser un blanc entre deux valeurs successives (type de format : 1 2 3 1). Trois possibilités sont envisagées pour les probabilités a priori des groupes : soit elles sont proportionnelles aux effectifs des groupes dans le fichier de données, soit elles sont égales, soit elles sont spécifiées par l'utilisateur. La validation de la règle de décision se fait soit à l'aide d'un échantillon test, tiré au hasard ou spécifié par l'utilisateur, soit par validation croisée uniquement pour DIPIND et DISLOG. Cette dernière possibilité est conseillée surtout si la taille de l'échantillon est très petite.

En sortie, on édite le tableau de classement issu du croisement de la partition à discriminer et la partition obtenue par la règle de décision, et le tout est stocké dans le fichier "résult". Ces sorties sont fournies pour l'échantillon d'apprentissage et, suivant les cas, pour l'échantillon test ou après évaluation par la validation croisée.

REFERENCES

[AND72] Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, **66**, 19-35.

[AND72] Anderson, J. A. (1975). Quadratic logistic discrimination. *Biometrika*, **62**, 149-154.

[BOC93] Bochi, S., Celeux, G. et Mkhadri, A. (1993). Le modèle d'indépendance conditionnelle : le programme DISIND. *La Revue Modulad*, n ?, 1-5.

[CEL94] Celeux, G., et Nakache, J. P. (1994). Analyse discriminante sur variables qualitatives. Paris: Polytechnica.

[CHO68] Chow, C. K., and Liu, C. N. (1968). Approximating discrete probability distributions with dependence tree. *IEEE Trans. Inform. Theory*, **IT-14**, 462-467.

[DAY67] Day, N. E., and Kerridge, D.F. (1967). A general maximum likelihood discriminant. *Biometrics*, **23**, 313-324.

[GEI66] Geisser, S. (1966). Predictive discrimination. In *Proc. of International Symposium on Multivariate Analysis*. P.R. Krishnaiah (Ed.), pp. 149-163, NY, Acad. Press.

[GEI93] Geisser, S. (1993). *Predictive Inference. An introduction*. Chapman and Hall.

[GOV90] Govaert, G. (1990). Classification binaire et modèles. *Revue de Stat. appl.*, *38*, N4, 43-58.

[GYL97] Gyllenberd and Koski. (1997). Posterior predictive distributions and maximal predictiveness for classification of multivariate binary data. *Technical Reports*, Department of Mathematiques, Royal Institute of Technology, Stockholm.

[HOS89] Hosmer, D. W., and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

[JOH96] Johnson, R.A., and Mouhab, A. (1996). A bayesien decision theory approach to classification problems. *J. Multivariate Analysis*, **vol. 56**, N2, 232-244.

[MKH94] Mkhadri, A., and Bochi, S. (1994). Regularized discrete probability distribution with dependence tree. *Rapport de Recherche INRIA*, n 2210.

[MKH97] Mkhadri, A., Celeux, C., and Nasroallah, A. (1997). Regularization in discriminant analysis: An Overview . *Computational Statistical \& Data Analysis*, **vol 23**, 403-423.

[NAD93] Nadif, M. et Marchetti, F. (1993). Classification de données qualitatives et modèles. *Revue de Stat. Appl.*, **41**, 55-59.

[NAS98] Nasroallah, A. (1998). Thèse d'Etat, Université Cadi Ayyad, Marrakech.

[WON77] Wong, A. C. K., and Wang, C. C. (1977). Classification of discrete biomedical data with error probability minimax. In *Proc. Seventh Inter. Conf. Cybern. Soc.*, Washington, DC, pp. 19-21.

(WON89] Wong, S. K. M., and Poon, F. C. S. (1989). Comments on approximating discrete probability distributions with dependence trees. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, **11**, No 3, 333-.