Le modèle d'indépendance conditionnelle: le programme DISIND

Sami Bochi, Gilles Celeux et Abdallah Mkhadri
INRIA Rocquencourt F78153 Le Chesnay

Résumé

On présente le modèle d'indépendance conditionnelle. On montre qu'il s'agit d'une méthode parcimonieuse et de référence pour l'analyse discriminante sur variables qualitatives, si l'on privilégie le point de vue décisionnel. On introduit le programme DISIND qui le réalise et qui valide la règle de décision par validation croisée.

Mots clés : *analyse discriminante, variables qualitatives, indépendance conditionnelle.*

1 Introduction

L'analyse discriminante décisionnelle a pour but de construire une règle de décision permettant d'affecter un individu décrit par p variables à l'un des k groupes $G_1, \ldots, G_k$ d'une partition définie a priori sur la population étudiée. Pour construire une telle règle de décision, on dispose d'un échantillon *d'apprentissage* de la population pour lequel l'affectation de chaque point à l'un des groupes $G_1, \ldots, G_k$ est connu. Le modèle statistique le plus général pour définir une règle de décision optimale est le modèle bayésien. La règle de décision bayésienne est celle qui minimise l'espérance du coût de mauvaise classification. Cette règle optimale dépend essentiellement des probabilités a priori $\pi_\ell, \ell = 1, \ldots, k$ des groupes qui vérifient

$$\pi_\ell \geq 0 \quad \text{pour tout } \ell \text{ et } \sum_{\ell=1}^{k} \pi_{\ell=1} \tag{1}$$

et des densités de probabilité par groupe $f_\ell(\mathbf{x}), \ell = 1, \ldots, k, \mathbf{x}$ appartenant à l'ensemble E des valeurs possibles des p variables descriptives. La règle bayésienne consiste à affecter tout vecteur $\mathbf{x}$ au groupe de probabilité a posteriori maximum

$$\mathbf{x} \quad \text{affecté à } G_r \Leftrightarrow r = arg \max_{s=1,\ldots,k} P(G_s/\mathbf{x}), \tag{2}$$

$P(G_s/\mathbf{x})$ désignant la probabilité a posteriori du groupe G_s. Par la formule de Bayes, on a

$$P(G_s/\mathbf{x}) = \frac{\pi_s f_s(\mathbf{x})}{\sum_{t=1}^{k} \pi_t f_t(\mathbf{x})} \tag{3}$$

La règle de Bayes peut donc s'écrire

$$\mathbf{x} \quad \text{affecté à } G_r \Leftrightarrow r = arg \max_{s=1,\ldots,k} \pi_s f_s(\mathbf{x}) \tag{4}$$

Ainsi, la construction effective d'une règle de décision revient à estimer les probabilités a priori $(\pi_\ell, \ell = 1, \ldots, k)$ et les densités de probabilité par groupe $(f_\ell(\mathbf{x}), \ell = 1, \ldots, k)$. L'opération principale de la discrimination à but décisionnnel est l'estimation des densités par groupe $f_s(\mathbf{x}), s = 1, \ldots, k$. Il faut, d'abord, faire des hypothèses sur la forme des densités $f_s(\mathbf{x})$. Ici, nous nous intéressons au cas où les p variables sont qualitatives. Pous simplifier les formules nous supposerons que les p variables sont binaires et donc l'espace $E = \{0,1\}^p$. Ainsi, il est naturel de supposer que les densités par groupe $f_s(x)$ sont des lois multinomiales sur l'ensemble des états possibles. Les lois multinomiales à considérer ont 2^p états possibles différents. Sous cette hypothèse, les paramètres des lois multinomiales sont bien sûr estimés par les fréquences observées qui sont les estimateurs du maximum de vraisemblance. Si $(x_1, \ldots, x_n)$ est l'échantillon d'apprentissage, utilisé pour construire la règle de décision, les estimations $\hat{f}_s(x)$ des densités $f_s(x)$ sont obtenues pour tout $s = 1, \ldots, k$ et pour tout $\mathbf{x}$ de $\{0,1\}^p$ par

$$\hat{f}_s(\mathbf{x}) = \frac{\sharp\{\mathbf{x}_i \in G_s \text{ et } \mathbf{x}_i = \mathbf{x}\}}{n_s} \tag{5}$$

où $n_s = \sharp\{x_i \in G_s\}$ ($\sharp$ désigne le cardinal d'un ensemble.)

D'après la formule (5), la règle de décision fondée sur le modèle multinomial complet est particulièrement facile à construire. Le défaut majeur du modèle multinomial complet est qu'il exige, pour chaque groupe a priori, l'estimation de $(2^p - 1)$ paramètres. Ce nombre devient très vite grand (si $p = 10, 2^p = 1024$). Ainsi, dans la plupart des problèmes pratiques, une estimation fiable des paramètres de ce modèle réclame des tailles d'échantillons énormes dont on ne peut en général pas disposer. On peut bien entendu se dire qu'une estimation très précise des lois $f_s(x)$ n'est pas nécessaire pour autant que les fréquences observées nous conduisent à la bonne décision. Le problème est en revanche plus aigu lorsque la fréquence d'un état est nulle dans chaque groupe. Le modèle d'indépendance conditionnnelle vise à résoudre ce problème d'identification par la diminution du nombre de paramètres à estimer.

2 Le modèle d'indépendance conditionnelle

Ce modèle suppose que les p variables binaires sont indépendantes dans chaque groupe. On dit, dans ce cas, que les variables sont *conditionnellement indépendantes*. Ce modèle signifie que les variables sont dépendantes, mais que la dépendance entre variables est entièrement expliquée par la connaissance des groupes a priori. Cet état de fait explique que l'on désigne aussi ce modèle sous le nom de modèle d'indépendance d'ordre *un* (Golstein, Dillon 1978) : si les variables étaient indépendantes, on parlerait de modèle d'indépendance d'ordre *zéro*.

Par ce modèle, les estimations $\hat{f}_s(x)$ des densités par groupe sont données par, pour tout $s = 1, \ldots, k$ et pour tout $\mathbf{x}$ de $\{0,1\}^p$

$$\hat{f}_s(\mathbf{x}) = \prod_{j=1}^{p} \frac{\sharp\{\mathbf{x}_i \in G_s \text{ tels que } x_i^j = x^j\}}{n_s} \tag{6}$$

où x^j représente la $j^{ème}$ cordonnée du vecteur x de $\{0,1\}^p$.

Le modèle d'indépendance conditionnelle appelle les commentaires suivants.

- Son grand intérêt est de proposer un nombre réduit de paramètres à estimer pour chaque groupe a priori : p au lieu de $2^p - 1$ pour le modèle multinomial complet. Cela étant, l'hypothèse d'indépendance conditionnelle peut être dans certains cas considérée trop simplificatrice. La discrimination logistique (cf. Hosmer et Lemeshow 1989 ou Celeux et Nakache 1993) offre un compromis entre l'éventuel trop grande simplicité du modèle d'indépendance conditionnelle et la trop grande complexité du modèle multinomial complet.

- A notre sens, un des grands atouts du modèle d'indépendance conditionnelle est de pouvoir offrir une méthode de sélection de variables

optimal en parfait accord avec le type de règle de décision construite. Il suffit, en effet, de sélectionner les variables indépendamment les unes des autres, ce qui n'induit aucune difficulté particulière. Les problèmes ne surgissent que lorsque qu'il faut sélectionner les variables conjointement.

- Cette méthode donne des résultats très satisfaisants dans beaucoup de situations. L'expérience montre que l'hypothèse d'indépendance conditionnelle est assez robuste. Si l'on pense à la facilité de sélectionner les variables pour ce modèle, on peut affirmer que cette méthode est une méthode d'analyse discriminante sur variables qualitatives de référence au même titre que l'analyse discriminante linéaire pour la discrimination sur variables quantitatives.
- Par des calculs simples (Celeux et Nakache 1993), on peut montrer que la règle de décision du modèle d'indépendance conditionnelle est une règle *linéaire* différente, en général, de la règle de la discrimination linéaire de Fisher, qui est réalisé par le programme DISC de Modulad. En pratique, on a constaté que ces deux méthodes donnaient souvent des règles de décision assez proches quant à leurs performances.

3 Le programme DISIND

Le programme DISIND réalise le modèle d'indépendance conditionnelle en privilégiant le point de vue décisionnel.

En entrée, les variables explicatives sont qualitatives avec un nombre quelconque de modalités. Trois possibilités sont envisagées pour les probabilités a priori des groupes : soit ellles 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. Si on opte pour l'utilisation d'un échantillon test tiré au hasard, le programme DISIND offre la possibilité de tirer plusieurs échantillons tests ce qui permet d'évaluer l'écart-type de l'erreur de classement. La procédure de validation croisée est conseillée.

En sortie, on édite le tableau de classement issu du croisement de la partition à discriminer et de la partition obtenue par la règle de décision. On peut si on le désire obtenir les probabilités d'affectation des individus aux groupes. Ces sorties sont, bien sûr fournies pour l'échantillon d'apprentissage et, suivant les cas, pour l'échantillon test ou après évaluation par la validation croisée.

Malgré son faible nombre de paramètres à estimer, il arrive que pour de petits échantillons le modèle d'indépendance conditionnelle conduise par la formule (6) à des estimations de probabilité par groupe nulles pour tous les groupes. Pour éviter ce problème et suivant les recommandations de

Titterington *et al.* 1981, le programme DISIND utilise une variante de la formule (6) qui lisse les fréquences marginales

$$\hat{f}_s(\mathbf{x}) = \prod_{j=1}^{p} \frac{\sharp\{x_i \in G_s \text{ tels que } x_i^j = x^j\} + \varepsilon^j}{n_s + 1} \tag{7}$$

avec $0 < \varepsilon^j < 1$.
En pratique, on prend $\varepsilon^j = 1/m_j$ où m_j est le nombre de modalités de la variable j.

Remerciements. Les auteurs remercient Ludovic Lebart pour ses remarques et ses conseils. C'est lui qui, notamment, a suggéré la possibilité de répéter le tirage d'un échantillon test.

Bibliographie

[1] Celeux, G. et Nakache, J. P. (éditeurs) (1993) *Analyse discriminante sur variables qualitatives*. Dunod (à paraître).

[2] Golstein, M. et Dillon W. R. (1978) *Discrete Discriminant Analysis*. Wiley.

[3] Hosmer, D. W. et Lemeshow, A. (1989) *Applied Logistic Regression*. Wiley.

[4] Titterington, D. M., Murray G. D., Murray L. S., Spiegelhalter D. J., Skene A. M., Habbema J. D. et Gelpke G. J. (1981) Comparison of discrimination techniques applied to a computer data set of head injuried patients. *J.R.S.S., A*, **144**, 145-175.