Introduction à la discrimination ; problématiques et méthodes

Gilbert Saporta

Pour terminer cette présentation, nous tenons à indiquer que ce présent livre est l'émanation d'un cours Modulad sur la discrimination, qui s'est tenu à Strasbourg en Novembre 1989, et qui fut organisé de main de maître par Danièle Grangé et Hélène Bigot que je remercie chaleureusement ainsi que les auteurs des différents articles. J'ai le plaisir à associer à ces remerciements les membres du club Modulad. Enfin je remercie Martine Cornélis pour son aide matérielle et Michel Bernadou pour avoir accepté d'accueillir cet ouvrage collectif dans la collection Didactique de l'INRIA.

Rocquencourt, le 21 Octobre 1990

Gilles Celeux

CHAPITRE 1

INTRODUCTION A LA DISCRIMINATION : PROBLEMATIQUE et METHODES

G. Saporta

Les méthodes de discrimination s'appliquent à des observations multidimensionnelles réparties en plusieurs groupes (ou sous-populations) définis a priori.
On peut leur assigner deux objectifs complémentaires :
- étudier si les variables recueillies permettent de distinguer les groupes,
- fournir des règles de classement pour prédire l'appartenance des observations aux groupes.

Le premier objectif correspond grosso-modo aux méthodes dites géométriques qui se rattachent aux méthodes descriptives de l'analyse des données (composantes principales, correspondances) ; le deuxième à des méthodes probabilistes utilisant des modèles plus ou moins contraignants.
Les méthodes de discrimination utilisent donc tout l'arsenal des techniques de la statistique descriptive et inférentielle et sont liées à la régression multiple, l'analyse de la variance, la théorie de la décision bayesienne etc...
C'est dire la richesse de ce sujet qui a suscité une littérature abondante (plus de 1000 articles dans des revues), de nombreux logiciels, et une foule d'applications dans les domaines les plus divers.

1. DE LA CRANIOLOGIE AU PUBLIPOSTAGE : 60 ANS D'APPLICATIONS

Les premiers travaux sur l'analyse discriminante remontent aux années 20 et ont été inspirés par des études concernant la reconnaissance de races humaines à partir de mesures craniennes (voir S. Das Gupta 1973).

Il est intéressant de noter que l'indice de distance généralisé, connu sous le nom de distance de Mahalanobis a été introduit plus de 10 ans avant la fonction de classement de Fisher, à partir d'un "coefficient de similitude raciale" proposé par K. Pearson en 1921 et 1926. Comme dans beaucoup d'autres domaines de la statistique, il y eut une ère Pearsonienne avant l'ère Fisherienne.

Mahalanobis s'intéressant aux races du Bengale en 1927, nota D^2 la distance généralisée. Hotelling proposa en 1931 son test T^2 d'égalité de deux vecteurs de moyennes qui ne diffère de D^2 que par une constante.

C'est en 1936 que Fisher propose sa fonction discriminante comme la combinaison linéaire de p variables qui maximise l'écart réduit entre les moyennes de deux groupes. Puis Welch en 1939 et Wald en 1944 donnèrent la formulation bayesienne du classement entre deux populations, que Rao étendit dans une série d'articles à plusieurs populations de distributions connues. Citons encore T.W. Anderson, en 1951, qui traite le cas où les paramètres sont estimés.

Depuis ces temps anciens (!), les applications de l'analyse discriminante n'ont cessé. Citons entre autres : en médecine, l'aide au pronostic (classification selon l'issue probable d'une maladie), l'aide au diagnostic (classification dans une forme de maladie comme les différents types d'hépatites) en météorologie (prévision des avalanches), en banque et assurance (prévision des risques, crédit-scoring), en marketing direct (sélection d'adresses dans des fichiers pour optimiser les envois de propositions commerciales), en reconnaissance de la parole.

Dans tous les cas, les données sont du type suivant : connaissant des échantillons de $n_1, n_2, ..., n_g$ observations appartenant à g groupes bien définis, on cherche une règle permettant d'affecter des observations d'appartenance inconnue à l'un des groupes.

2. LES METHODES

Elles diffèrent selon la nature des variables et l'optique recherchée : analyse de la situation ou classement.

2.1 Méthodes géométriques

2.1.1 Variables numériques : analyse factorielle discriminante

L'analyse factorielle discriminante consiste dans la recherche de nouvelles variables combinaisons linéaires des p variables initiales qui fournissent les meilleures séparations des groupes au sens habituel de l'analyse de la dispersion : rapport dispersion inter / dispersion intra. Géométriquement cela revient à chercher des axes de $\mathbb{R}^p$ tels que les centres g_i des groupes s'y projettent avec une dispersion maximale, tout en minimisant en moyenne la dispersion de chaque groupe.

Si on note T, B, W respectivement les matrices de dispersion totale, inter et intra, les combinaisons linéaires discriminantes b qui réalisent des maximums successifs de

$$\frac{b'B\,b}{b'\,T\,b} \text{ ou de } \frac{b'\,B\,b}{b'\,W\,b}$$

sont les vecteurs propres de $W^{-1}\,B$ ou de $T^{-1}\,B$.

L'analyse factorielle discriminante apparait alors comme une analyse en composantes principales du nuage des centres de gravité des groupes avec la métrique T^{-1} ou la métrique W^{-1} (métrique de Mahalanobis). Le nombre d'axes discriminants, donc de variables discriminantes est en général égal à g-1 si $p > g$. Cette méthode est également appelée analyse canonique discriminante car on peut la retrouver en effectuant l'analyse canonique (recherche de combinaisons linéaires en corrélation maximale) des deux ensembles suivants de variables : d'une part les p variables numériques observées, d'autre part les g indicatrices des groupes.

Lorsque $g = 2$, il n'y a qu'une seule variable discriminante qui s'obtient aisément puisque l'axe discriminant est la droite des centres :

$b = W^{-1} (g_1 - g_2)$. On retrouve ainsi la fonction linéaire discriminante de Fisher.

Toujours dans le cas de 2 groupes, l'analyse discriminante se ramène formellement à une régression entre une variable y bivalente (valant par exemple 0 sur le groupe 1 et 1 sur le groupe 2) et l'ensemble des variables numériques, ce qui permet d'utiliser des logiciels de régression mais pas tous leurs résultats, en particulier les tests de signification, car les hypothèses de normalité des résidus du modèle linéaire ne sont pas vérifiées.

Les variables discriminantes étant définies à un coefficient multiplicatif près, il est d'usage de choisir celui-ci de telle sorte que l'on retrouve dans l'espace de projection les distances de Mahalanobis entre groupes par simple application du théorème de Pythagore.

2.1.2 Variables qualitatives

On ne peut évidemment trouver des combinaisons linéaires des variables, à moins d'en réaliser un codage particulier. La méthode Disqual proposée par l'auteur permet de se ramener au cas précédent en remplaçant les variables qualitatives par les composantes de l'analyse factorielle des correspondances multiples. Cette façon de faire est très utilisée pour la notation des demandeurs de crédit. Dans le cas de deux groupes, le programme DIS2G présent dans MODULAD réalise ces opérations.

2.1.3 Limitations

Les méthodes géométriques, dont l'objectif est de trouver des variables discriminantes et une représentation des observations mettant en évidence leur séparation en groupes, permettent également d'effectuer des classements en utilisant des règles simples basées sur la distance de Mahalanobis aux centres des groupes. Mais de tels procédés ne permettent pas de calculer des probabilités d'affectation et ne sont pas nécessairement optimaux au sens de la théorie de la décision. On recourt alors aux modèles probabilistes. Notons ici que certaines applications ne relèvent que de la description comme les célèbres Iris de Fisher.

2.2 Méthodes probabilistes

Elles reposent sur la règle bayesienne : on classe une observation x dans le groupe pour lequel la probabilité a posteriori est maximale. Si les p_i sont les probabilités a priori et les $f_i(x)$ les densités de probabilité pour chaque population, la probabilité a posteriori est :

$$P(G_i \mid x) = \frac{p_i f_i(x)}{\sum p_i f_i(x)} .$$

On cherchera donc le max de $p_i f_i (x)$.

Si la connaissance des probabilités a priori ne pose en général pas de problème, il n'en est pas de même de celles des $f_i(x)$.

2.2.1 Le modèle normal.

Tout naturellement les théoriciens se sont penchés sur le cas où les $f_i(x)$ sont des densités normales p - dimensionnelles $N_p(\mu_i ; \Sigma_i)$.

Si les matrices Σ_i sont différentes, la règle de Bayes conduit à des formules de classement quadratiques en x , si les Σ_i sont égaux, à des formules linéaires.

Si de plus les p_i sont égaux on retrouve alors la règle géométrique de classement au groupe le plus proche au sens de la distance de Mahalanobis.

En pratique tous les paramètres doivent être estimés, ce qui pose entre autres problèmes celui de décider si les Σ_i sont égaux ou non. Le fait d'utiliser des estimations séparées des Σ_i , donc des formules quadratiques, peut conduire à des gains illusoires vu le nombre de paramètres à estimer. Pour deux groupes il est possible de rechercher la meilleure formule linéaire avec $\Sigma_1 \neq \Sigma_2$ (méthode d'Anderson Bahadur).

2.2.2 La discrimination logistique

Il s'agit d'une approche que l'on peut qualifier de semi-paramétrique où l'on modélise la probabilité a posteriori et non les densités pour chaque groupe.

On pose $P(G/x) = \frac{\exp(\theta_0 + \theta' x)}{1 + \exp(\theta_0 + \theta' x)}$.

Les paramètres θ_0 et θ sont alors estimés par une méthode de maximum de vraisemblance conditionnel. Cette méthode qui a la faveur des économètres, contient comme cas particuliers le modèle normal et aussi certains modèles log-linéaires. Bien sûr, si le modèle normal est vrai, il vaut mieux utiliser la théorie normale qui correspond à des estimations du M.V. non conditionnelles.

2.2.3 Méthodes non paramétriques

Elles consistent à estimer soit les densités $f_i(x)$ soit la probabilité a posteriori $P(G_i/x)$. Le premier cas peut être traité à l'aide de toute technique d'estimation de densité ; la plus connue étant la méthode du noyau :

$$f_i(x) = \frac{1}{n_i} \sum_{x_j \in G_i} K\left(\frac{x - x_j}{\mu}\right) .$$

La méthode des k-plus proches voisins est une technique très simple d'estimation de la probabilité a posteriori, surtout utilisée pour classer directement x : on cherche les k voisins les plus proches de x au sens d'une certaine métrique et on classe dans le groupe majoritaire.
Ces méthodes ont l'inconvénient vis à vis des précédentes d'être des boites noires où l'on ne peut évaluer l'influence des diverses variables.

2.2.4 Méthodes probabilistes pour variables qualitatives

Si les variables explicatives $X_1, X_2, \ldots, X_p$ sont qualitatives à respectivement $m_1, m_2, \ldots, m_p$ catégories, il suffirait en théorie pour prédire l'appartenance d'une observation x à un groupe de considérer la fréquence de ce groupe conditionnellement aux X_i . C'est le modèle multinomial.

Le produit $\prod_{i=1}^{p} m_i$ étant en général très grand, cette approche est irréaliste et il faut utiliser des modèles simplificateurs exprimant $P(G_j/x)$ à l'aide d'un petit nombre de paramètres. On citera ici la régression qualitative de Daudin (modèle linéaire sur les probabilités) et le modèle log-linéaire sur les rapports de probabilité. D'autres modèles ont été proposés (Lancaster) mais n'ont pas été réellement appliqués.

2.4 Discrimination par arbre

Les arbres de segmentation proposent une approche toute différente. On détermine pour chaque variable la dichotomie optimale au sens d'un certain critère lié à la séparation de la population en g groupes et on choisit la meilleure variable pour effectuer une dichotomie de la population. On recommence alors dans chacun des 2 groupes etc...

Ces méthodes très utilisées en marketing ne jouissaient plus de la considération des statisticiens (combinatoire rebutante, absence d'estimation des probabilités d'erreur ...) jusqu'à ces dernières années avec la parution des travaux de L. Breiman et J. Friedman qui utilisent un nouveau critère dit d'impureté, lié au mélange des sous-populations à un noeud de l'arbre et une procédure de validation croisée à l'aide d'un échantillon test pour ne conserver que les branches utiles de l'arbre.
Cette technique conduisant à un raisonnement séquentiel proche de celui des utilisateurs qui hiérarchisent les variables connait un regain d'intérêt très net.

3. LES PROBLEMES

Dans l'application des méthodes de discrimination, le praticien se trouve fréquemment confronté à divers problèmes assez différents les uns des autres.

3.1 Les mélanges de variables

Il est courant d'avoir à la fois des prédicteurs numériques et qualitatifs. Certaines méthodes permettent de les traiter simultanément comme la discrimination logistique mais pas d'autres. On peut alors recourir à des recodages, le procédé le plus fréquent consistant à rendre qualitatives les variables par un découpage en classes, ce qui permet d'ailleurs de pouvoir exploiter des non linéarités éventuelles.

Dans le cas de deux groupes, notons une possibilité peu utilisée : celle du modèle linéaire général puisque la discrimination linéaire peut s'effectuer au moyen d'une régression multiple : il suffit ici d'introduire les indicatrices des catégories des variables qualitatives comme prédicteurs.

3.2 Mesures de qualité et validation

Les indicateurs statistiques tels le D^2 de Mahalanobis, la Trace de Pillai (somme des valeurs propres de $V^{-1} B$), le Λ de Wilks etc... n'intéressent pas toujours le praticien qui veut connaître les performances en termes de taux d'erreur de classement. Ces taux d'erreurs ne sont pas liés de façon évidente aux statistiques précitées dans le cas gaussien et encore moins dans les autres cas.

On mesure les taux d'erreurs en appliquant les règles de classement à des observations de provenance connue.

La méthode de resubstitution qui consiste à réutiliser les mêmes données que celles qui ont servi à bâtir les règles, présente un biais évident. Il faut utiliser de préférence un échantillon test d'observations supplémentaires. Lorsque l'on ne dispose pas de suffisamment de données il faut alors recourir à des méthodes de validation croisée (on reclasse chaque observation à l'aide de n-1 autres) ou de rééchantillonnage bootstrap.

3.3 Choix de variables

Un grand nombre de motifs peuvent conduire à restreindre le nombre de prédicteurs : moindre coût, meilleure stabilité des résultats entre autres.
Comme dans d'autres domaines il faut définir des critères de choix et de procédures de sélection.

En dépit de ce qui a été signalé au paragraphe précédent, la plupart du temps on compare deux ensembles de variables en termes d'indicateurs statistiques comme le Λ de Wilks ou le D^2 de Mahalanobis pour 2 groupes et non selon les performances réelles : l'avantage en est une simplification des calculs car tout se déduit des matrices de covariance et il n'est pas nécessaire de procéder au classement des observations à chaque essai.

Comme l'exploration exhaustive des sous-ensembles de variables est souvent impossible on utilise des méthodes pas à pas. Notons cependant que pour deux groupes, l'identité avec une régression permet d'utiliser des méthodes très performantes de recherche exhaustive comme celle de Furnival et Wilson.

La méthode la plus courante consiste à augmenter progressivement le nombre de variables jusqu'à ce que l'on ne puisse plus rajouter ni enlever de prédicteurs selon un test de type F partiel.

L'ajout d'une variable à un sous-ensemble ayant pour effet d'augmenter le D^2 et de diminuer le Λ, on cherche à maximiser la variation à chaque pas.