

STATISTIQUE ET ANALYSE DES DONNÉES

P. BAUFAYS

J.-P. RASSON

Une nouvelle règle de classement, utilisant l'enveloppe convexe et la mesure de Lebesgue

Statistique et analyse des données, tome 9, n° 2 (1984), p. 31-47

http://www.numdam.org/item?id=SAD_1984__9_2_31_0

© Association pour la statistique et ses utilisations, 1984, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

UNE NOUVELLE REGLE DE CLASSEMENT, UTILISANT
L'ENVELOPPE CONVEXE ET LA MESURE DE LEBESGUE

P. BAUFAYS et J.-P. RASSON

Unité de Statistique - Département de Mathématique
Facultés Universitaires N.D. de la Paix
Rempart de la Vierge 8 - B-5000 NAMUR, BELGIQUE

Résumé : Nous proposons une nouvelle méthode d'analyse discriminante associée au modèle statistique utilisé en classification automatique par Hardy et Rasson (1982). Cette règle est la suivante : nous calculons la différence entre les mesures de Lebesgue des enveloppes convexes de l'échantillon étiqueté de chaque groupe avec et sans l'individu à classer; nous l'affectons au groupe pour lequel cette différence est minimale. Lorsqu'un individu appartient à plus d'une enveloppe convexe, aucune décision n'est prise de son sujet. La courbe de décision associée à cette règle est linéaire par morceaux; nous indiquons une procédure pour la calculer et nous l'illustrons sur différents exemples. Enfin, nous décrivons la construction d'un ensemble de références pour lequel la règle du plus proche voisin est identique à la règle originale appliquée à l'échantillon original.

Mots clés : Analyse discriminante paramétrique, Courbe de décision linéaire par morceaux, Enveloppe convexe, Mesure de Lebesgue, Règle du plus proche voisin.

Abstract : We propose a new method of discriminant analysis, associated with the statistical model used in cluster analysis by Hardy and Rasson (1982). This rule is the following : we compute the difference between the Lebesgue measures of the convex hulls of each group labelled sample with and without the individual to be assigned; this is allocated to the population with the

smallest difference. No decision is taken about individuals belonging or none than one convex hull. The decision surface is piecewise linear; we indicate a method to compute this and we illustrate it on several examples. Finally, we describe the construction of a reference set for which the nearest neighbor rule is the same as the new method applied to the original sample.

Keywords : Convex hull, Lebesgue measure, Parametric discriminant analysis, Piecewise linear discriminant curve, Nearest neighbor rule.

1 - INTRODUCTION

L'analyse discriminante vise à déterminer des règles optimales de classement d'un individu dans un parmi plusieurs groupes (k); la seule information à leur propos est un échantillon d'individus dont l'affectation est connue. Pour chaque individu n variables ont été observées. Le problème est alors celui de déterminer une partition de R^m en k sous-ensembles de telle sorte que tout individu appartenant à un sous-ensemble de la partition soit affecté au groupe correspondant à celui-ci.

De nombreuses méthodes ont été proposées dans la littérature (Andrews (1972), Duda et Hart (1973), Fard (1981)). Certaines - déterministes - imposent la forme paramétrique de la frontière entre les ensembles d'affectation à des groupes différents. Cette surface est déterminée de telle sorte que le nombre d'individus mal classés de l'échantillon étiqueté soit minimal. Les autres méthodes - statistiques - nécessitent la connaissance des probabilités a priori d'appartenance aux différents groupes; elles visent à minimiser le risque bayésien. Suivant qu'aucune hypothèse n'est formulée à propos des distributions, on distingue les méthodes non paramétriques (k -NN, estimation de densités, ...) et paramétriques, où l'on suppose généralement que les distributions sont normales. C'est dans cette dernière catégorie que se situe la méthode décrite ci-après. Nous considérons, en effet, la règle de classement associée à un modèle statistique nouveau, proposé et utilisé en classification automatique par Hardy et Rassen (1982). Les résultats intéressants tant théoriques que pratiques, fournis par cette méthode constituent un encouragement certain à considérer le même modèle en analyse

discriminante.

2 - LE MODELE - LE CRITERE

Nous rappelons le modèle proposé par Hardy et Passon et nous renvoyons le lecteur à cette étude pour une description et une justification plus complète.

Ce modèle suppose que tout individu est uniformément distribué dans l'union C de k domaines convexes disjoints (C_1, \dots, C_k) , ce qui signifie que l'échantillon étiqueté est une réalisation d'un processus de Poisson dans C . Chaque domaine correspond à un groupe. La vraisemblance est donc égale à :

$$f(x) = (1 / \sum_{i=1}^k m(C_i)) \cdot \sum_{i=1}^k I_{C_i}(x)$$

où x est le vecteur des variables observées pour un individu, $m(\cdot)$ désigne la mesure de Lebesgue et $I_A(\cdot)$ la fonction indicatrice de l'ensemble A . En posant :

$$p_i = m(C_i) / \sum_{j=1}^k (m(C_j)) ,$$

la vraisemblance s'écrit encore :

$$f(x) = \sum_{i=1}^k (p_i / m(C_i)) I_{C_i}(x)$$

et nous voyons que $f(x)$ est le mélange de k densités uniformes (sur chacun des domaines C_i) avec des poids p_i proportionnels aux mesures des domaines.

Si nous supposons tous les coûts de mauvaise classification égaux, la règle qui minimise le risque bayésien affecte l'individu x au groupe i ssi :

$$p_i f_i(x) \geq p_j f_j(x) \quad , \quad j=1, \dots, k, \quad j \neq i .$$

Les paramètres des distributions - les domaines C_i - sont inconnus. Nous estimons C_i par maximum de vraisemblance, c'est-à-dire par l'enveloppe convexe $H(X_i)$ des individus étiquetés X_i du groupe i . Un traitement complet de l'estimation d'un domaine convexe borné dont on connaît un

échantillon uniformément distribué est proposé par Rason (1979).

Tout individu appartenant à $H(X_i)$ est donc affecté au groupe i . Supposons maintenant que x n'appartient à aucune enveloppe convexe. S'il était affecté au groupe i , nous aurions :

$$p_i f_i(x) = 1 / \left(\sum_{j=1}^k m(H(X_j)) + m(H(X_i \cup \{x\})) - m(H(X_i)) \right) ;$$

$$p_j f_j(x) = 0 \quad , \quad j=1, \dots, k, \quad j \neq i .$$

Définissons :

$$S_i(x) = m(H(X_i \cup \{x\})) - m(H(X_i)) .$$

Dès lors :

$$p_i f_i(x) = 1 / \left(\sum_{j=1}^k m(H(X_j)) + S_i(x) \right) .$$

D'autre part, puisque les domaines sont supposés disjoints, nous ne pouvons affecter x au groupe i que si $H(X_i \cup \{x\})$ et $H(X_j)$ sont disjoints pour tout $j \neq i$. Dans le cas contraire, nous posons arbitrairement :

$$S_i(x) = +\infty .$$

$S_i(x)$ peut être interprété comme une mesure de la perte de confiance du groupe i lorsqu'on lui adjoit l'individu x . Et le critère d'affectation est :

"classer x dans le groupe i ssi

$$S_i(x) < S_j(x) \quad , \quad j=1, \dots, k, \quad j \neq i ."$$

Nous ne supposons plus maintenant que les domaines C_i sont disjoints, afin de rendre notre règle applicable dans toutes les situations. Dans ce cas, un individu qui appartient à une enveloppe convexe est affecté au groupe correspondant. S'il appartient à plusieurs enveloppes convexes, nous ne prenons aucune décision à son sujet. Enfin, s'il n'appartient à aucune enveloppe convexe, il sera affecté au groupe i ssi :

$$S_i(x) < S_j(x) \quad , \quad j=1, \dots, k, \quad j \neq i .$$

La non affectation lorsqu'un individu appartient à plusieurs enveloppes convexes correspond exactement à la discrimination partielle proposée par Chow (1970) et Habbema et al. (1974). Supposons, en effet,

que l'individu x appartient aux enveloppes convexes $H(X_1)$, ... , $H(X_p)$, et n'appartient pas à $H(X_{p+1})$, ... , $H(X_k)$. La probabilité a posteriori que x appartienne au groupe i est alors égale à $1/p$ ($1 \leq i \leq p$) et à 0 ($p+1 \leq i \leq k$). En effet, cette probabilité a posteriori est égale à $p_i f_i(x) / (p_1 f_1(x) + \dots + p_k f_k(x))$. Or, dans ce cas :

$$p_i f_i(x) = 1 / \left(\sum_{j=1}^k m(H(X_j)) \right) , \quad i=1, \dots, p ,$$

$$= 0 , \quad \text{si non.}$$

La non affectation des cas les plus coûteux a été introduite dans de nombreuses règles afin de diminuer les erreurs de classement. Quelques exemples sont cités dans la bibliographie (Chow (1970), Degtyar et Finkelshteyn (1974), Habbema et al. (1974), Kendall (1979), Tomak (1976), Wilcox et Wilson (1980)).

La méthode proposée ici étend et justifie la méthode empirique proposée à la fois par Kendall (1979) et Degtyar et Finkelshteyn (1974); celle-ci et la nouvelle règle coïncident pour les individus appartenant à au moins une enveloppe convexe, mais la plus ancienne ne traitait pas les autres individus. Levrat (1979) suggère, lorsque les enveloppes convexes sont disjointes, d'affecter tout individu au groupe de l'enveloppe convexe la plus proche. Nous mettons ainsi en évidence un aspect original de notre critère : alors que la plupart des règles d'affectation sont définies par une distance, notre méthode généralise la distance sur la droite réelle en surface dans le plan, en volume dans l'espace usuel, en hypervolume dans les espaces à plus de trois dimensions.

On notera enfin que, sur la droite réelle, les règles du plus proche voisin (PPV) de Levrat et la nôtre coïncident lorsque les enveloppes convexes sont disjointes.

3 - AFFECTATION ET SURFACE DE DECISION

L'affectation d'un individu et le calcul de la surface de décision nécessitent l'évaluation de $S_1(x)$. Nous décrivons une procédure pour ce faire dans le plan; son principe peut être généralisé pour des espaces de dimension supérieure. Le développement du calcul géométrique sur ordinateur, domine en plein essor à l'heure actuelle (Toussaint (1980))

devrait fournir des outils qui permettent un calcul performant du classement et des surfaces de décision dans tout espace.

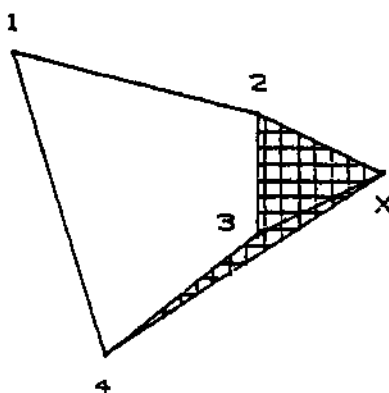
3.1 - Calcul de S_i

Nous connaissons $H(X_i)$ et voulons calculer $H(X_i U \{x\})$. $H(X_i)$ est borné par un nombre fini de droites (hyperplans dans l'espace) déterminées par ses faces. Chaque droite définit deux demi-espaces dont l'un contient X_i ; à chaque droite, nous associons l'entier j , égal à 1 si x et X_i sont dans le même demi-espace défini par celle-ci et 0 sinon. Il est évident que toute face de $H(X_i)$ pour laquelle $j=1$ sera aussi une face de $H(X_i U \{x\})$, au contraire de toute face pour laquelle $j=0$; il nous reste donc à déterminer les faces de $H(X_i U \{x\})$ contenant x . Tout(e) sommet (arête) de $H(X_i)$ est l'intersection de deux faces; si pour l'une d'elles $j=0$, et pour l'autre $j=1$, ce(tte) sommet (arête) et x définissent une face de $H(X_i U \{x\})$. Nous appelons de tel(le)s sommets (arêtes), sommets (arêtes) de tangence de x .

$S_i(x)$ est égale à la somme des mesures des p triangles (pyramides) ayant x pour sommet et pour base une face G_i ($i=1, \dots, p$) pour laquelle $j=0$ (voir figure 3-1). Soit n_i la normale unitaire à G_i pointant vers l'extérieur de $H(X_i)$, et soit a_i un point de cette face. Nous avons, à une constante multiplicative près, qui dépend de la dimension de l'espace :

$$S_i(x) = \sum_{j=1}^p g_j \langle x - a_j, n_j \rangle, \quad (1)$$

où g_j est la mesure de la face G_j . D'autre part, si tous les entiers j relatifs à x sont égaux à 1, x appartient à $H(X_i)$, et dès lors $S_i(x) = 0$. Pour les droites déterminées par les couples de sommets (2,3) et (3,4) $j = 0$, et pour celles par (1,2) et (1,4) $j = 1$. Les sommets de tangence sont les sommets 2 et 4. $S_i(x)$ est la somme des surfaces des triangles (2,3,x) et (3,4,x) (en hachuré).

Figure 3-1. Calcul de S_1 .

Lorsque les domaines sont supposés disjoints, il faut en plus vérifier que $H(X_i U \{x\})$ est disjoint de tout $H(X_j)$, $j=1, \dots, k$, $j \neq i$. Pour vérifier le caractère disjoint de deux enveloppes convexes, la littérature suggère deux solutions. La première consiste à calculer explicitement l'intersection de $H(X_j)$ et de $H(X_i U \{x\})$: des algorithmes efficaces ont été proposés pour R^2 par Shamos (1975) et pour R^2 et R^3 par Muller et Preparata (1978). La seconde ne nécessite pas le calcul de $H(X_i U \{x\})$; elle consiste à appliquer l'algorithme du Perceptron (Andrews (1972)) aux ensembles F_i - constitué des sommets de $H(X_i)$ et de x - et E_j , contenant les sommets de $H(X_j)$. Elle est applicable à tous les espaces.

Une troisième solution permet aussi de vérifier le caractère disjoint sans connaître $H(X_i U \{x\})$, en calculant les zones d'ombre. La zone d'ombre associée à C_1 est l'ensemble des points de l'espace qui ne peuvent être affectés à C_2 parce que $H(X_2 U \{x\})$ et $H(X_1)$ ne sont pas disjoints. En optique géométrique, si $H(X_1)$ était un corps opaque et $H(X_2 U \{x\})$ une source lumineuse, certains rayons de lumière issus de $H(X_2 U \{x\})$ ne pourraient atteindre un point de la zone d'ombre parce que $H(X_1)$ forme un écran. Les zones d'ombre sont déterminées par les tangentes communes aux deux enveloppes convexes telles que chaque demi-espace qu'elles définissent contient un des deux groupes (voir figure 3-2). Les zones d'ombre sont déterminées par les tangentes communes à $H(X_1)$ et à $H(X_2)$, qui séparent linéairement ces deux ensembles.

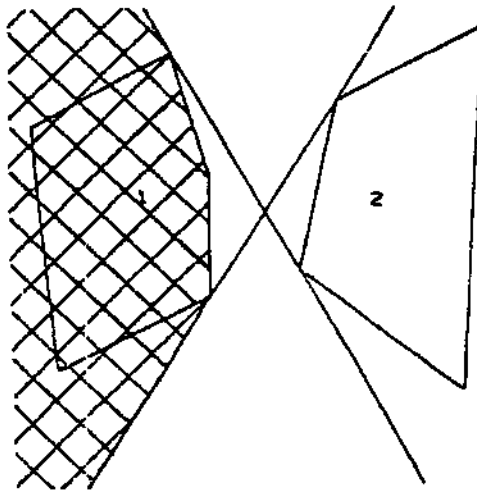


Figure 3-2. Zone d'ombre relative à C_1 en hachuré et tangentes séparatrices.

3.2 - Courbe de décision

La frontière entre les sous-ensembles d'affectation à deux groupes - ou courbe de décision - est l'ensemble :

$$\{x : S_1(x) = S_2(x)\} .$$

Nous montrons que cette courbe est linéaire par morceaux. Soit x_0 tel que $S_1(x_0) = S_2(x_0)$. Notons F_1^i, \dots, F_p^i (resp. G_1^i, \dots, G_p^i) les faces de $H(X_i)$ contenant une arête de tangence, pour lesquelles $j=1$ (resp. $j=0$). $D_i(x_0)$ est l'ensemble des points possédant la même structure de tangence à $H(X_i)$ que x_0 . Pour tous les points de cet ensemble, les mêmes entiers j sont associés à chaque face de $H(X_i)$:

$$D_i(x_0) = F_1^{i+} \cap \dots \cap F_p^{i+} \cap G_1^{i-} \cap \dots \cap G_p^{i-}$$

où F^* désigne le demi-espace défini par la face F contenant le groupe i . $D_i(x_0)$ est un ensemble convexe non vide. Un domaine $D_i(x_0)$ est illustré à la figure 3-3.

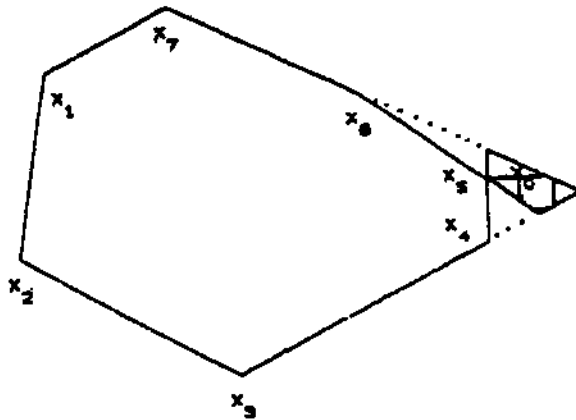


Figure 3-3. Tous les points du domaine hachuré possèdent la même structure de tangence.

De plus la formule (1) est valable pour tout point de $D_1(x_0)$. Dès lors, tous les points de $D_1(x_0) \cap D_2(x_0)$ satisfaisant l'équation suivante appartiennent à la courbe de décision :

$$\sum_{j=1}^{P_1} g_j^1 \langle x - a_j^1, n_j^1 \rangle = \sum_{j=1}^{P_2} g_j^2 \langle x - a_j^2, n_j^2 \rangle \quad (2)$$

où l'indice supérieur (i) désigne le groupe.

3.3 - Calcul de la courbe de décision

Nous proposons ici une méthode de calcul de la courbe de décision, que nous détaillons dans le cas du plan.

En premier lieu, nous déterminons un point x_0 de la courbe de décision. Si les deux enveloppes convexes ont une intersection non vide, ce sera n'importe quel point commun aux deux frontières. Dans le cas contraire, nous cherchons une racine de $S_1(x) - S_2(x)$, par l'algorithme dichotomique, sur un segment de droite dont chaque extrémité est un sommet de l'une des enveloppes convexes.

Nous calculons alors $D_1(x_0)$ et $D_2(x_0)$ et la droite qui supporte la partie de la courbe de décision dans $D_1(x_0) \cap D_2(x_0)$, suivant (2). Soient x_1 et z_1 les extrémités du segment inclus dans $P_1 =$

$D_1(x_0) \cap D_2(x_0)$. Chacun appartient à une droite définie par une face d'un des convexes. Dès lors, ils appartiennent chacun à deux ensembles "du même type" que P_1 (voir figure 3-4). Nous appellerons dans la suite de tels ensembles, pavés élémentaires, parce que chacun contient une partie élémentaire (un segment de droite) de la courbe de décision. Connaissant P_1 , il est aisé de calculer le second pavé élémentaire P_2 .

Sans perdre de généralité, nous supposons que x_1 appartient à une droite définie par une face du premier groupe. Soient i_1 et i_2 les indices des sommets de tangence de x_0 à C_1 , tels que les sommets $i_1+1, i_1+2, \dots, i_2-1$ ne soient pas sommets de $H(x_1 \cup \{x_0\})$.

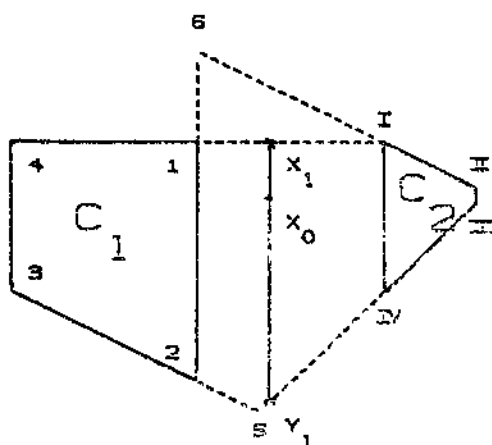


Figure 3-4. Calcul de la courbe de décision.

Le point initial x_0 est marqué par une croix. Le premier pavé élémentaire est limité par les segments $(1,2)$, $(2,5)$, $(5,IV)$, (IV,I) , $(I,1)$. Les sommets de tangence de x_0 relatifs au premier (resp. second) convexe sont 1 et 2 (resp. I et IV). Les points x_1 et y_1 sont notés par *. Le second pavé auquel x_1 appartient est le triangle de sommets 5, I et 1. Dans ce pavé, les sommets de tangence au groupe 1 (resp. 2) sont 2 et 4 (resp. I et IV). x_1 appartient à la face définie par les sommets 1 et 4.

Définissons les droites D_1, D_2, D_3 et D_4 déterminées par les sommets (respectivement) i_1 et i_1-1 (D_1), i_1 et i_1+1 (D_2), i_2-1 et i_2 (D_3) et i_2 et i_2+1 (D_4). Certaines de ces droites

peuvent être confondues. Nous avons

$$D_1(x_0) = D_1^+ \cap D_2^- \cap D_3^- \cap D_4^+.$$

x_1 appartient à l'une de ces droites :

- x_1 appartient à D_1 ; dans ce cas, P_2 n'est pas dans le même demi-espace défini par D_1 que P_1 . Aussi, l'indice j associé à D_1 pour les points de P_2 est 0. Dès lors, les sommets de tangence au premier convexe (pour les points de P_2) sont i_1-1 et i_2 ; soit D_5 la droite définie par les sommets i_1-2 et i_1-1 . Nous avons :

$$D_1(x_1) = D_5^+ \cap D_1^- \cap D_3^- \cap D_4^+.$$

- De la même façon : x_1 appartient à D_2 ; les sommets de tangence sont i_1+1 et i_2 ; D_5 est définie par i_1+1 et i_1+2 :

$$D_1(x_1) = D_2^+ \cap D_5^- \cap D_3^- \cap D_4^+.$$

- x_1 appartient à D_3 ; les sommets de tangence sont i_2-1 et i_2 ; D_5 est définie par i_2-2 et i_2-1 :

$$D_1(x_1) = D_1^+ \cap D_2^- \cap D_5^- \cap D_3^+.$$

- x_1 appartient à D_4 ; les sommets de tangence sont i_2+1 et i_2 ; D_5 est définie par i_2+2 et i_2+1 :

$$D_1(x_1) = D_1^+ \cap D_2^- \cap D_4^+ \cap D_5^-.$$

Nous avons alors :

$$P_2 = D_1(x_1) \cap D_2(x_0).$$

Nous calculons alors le segment de la courbe de décision dans ce pavé élémentaire. Il a pour extrémités x_1 et x_2 . Nous continuons ensuite le même procédé avec x_2 . Les pavés élémentaires sont en nombre fini ; de plus, tous ne contiennent pas un segment de la courbe de décision. Dès lors, après un nombre fini d'étapes, ou bien la méthode revient en z_1 , ou bien elle s'arrête lorsqu'elle obtient un pavé non borné dans lequel la courbe de décision est une demi-droite. Dans ce cas, nous recommençons le même procédé avec z_1 comme point initial.

Lorsque les domaines convexes sont supposés disjoints, nous devons, en

autre, nous assurer que la courbe de décision a une intersection vide avec chacune des zones d'ombre.

Dans les espaces de dimension supérieure à deux, le calcul de la surface de décision se fait par une méthode dont le principe est semblable à celui décrit ci-dessus. Les sommets de tangence deviennent des arêtes de tangence. La différence essentielle vient de ce que l'intersection de l'hyperplan définissant la surface de décision dans un pavé élémentaire avec ce pavé ne se résume plus à deux points; et le calcul de cette intersection devient le point délicat.

Notre critère d'affectation et le calcul de la surface de décision nécessitent le calcul des enveloppes convexes des groupes et des mesures de Lebesgue de polyèdres convexes. De nombreux algorithmes ont été proposés pour le calcul de l'enveloppe convexe d'un nombre fini de points et ce sujet continue à constituer la préoccupation de nombreux chercheurs (Toussaint (1980)). D'autre part, l'algorithme le plus performant de calcul de la mesure d'un polyèdre convexe est, à notre connaissance, celui de Lasserre (1984).

De nombreuses simulations ont montré que le nombre de segments de droite qui forment la courbe de décision, dans le plan, est du même ordre de grandeur que le nombre total d'individus étiquetés, qui sont sommets d'une enveloppe convexe.

Nous présentons aux figures 3-5, 3-6 et 3-7 trois exemples de courbes de décision. Le premier discrimine un groupe de 11 hétérosexuels d'un groupe de 15 homosexuels sur base des taux d'étiolacholone et d'androstérone dans l'urine (Rand (1981)). Le deuxième illustre le cas où les enveloppes convexes ne sont pas disjointes en utilisant les célèbres données de Fisher (1936). Nous considérons uniquement les largeurs et les longueurs des pétales pour les variétés d'iris virginica et versicolor. Enfin, le troisième exemple traite une situation où le nombre de populations est supérieur à deux (Données de Ruspini (1959)).

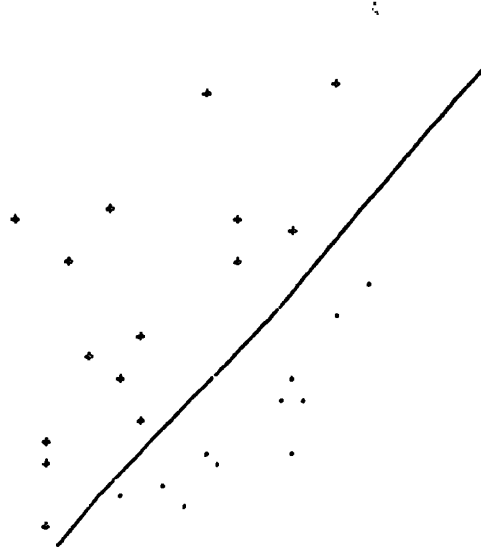


Figure 3-5. Homosexuels et hétérosexuels.

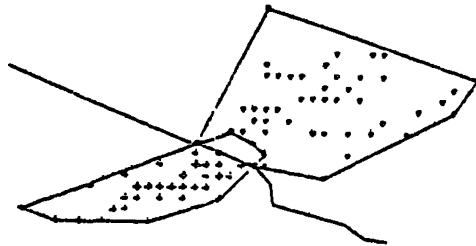


Figure 3-6. Iris.



Figure 3-7. Données de Ruspini.

1 - UN ENSEMBLE DE REFERENCE EQUIVALENT

Parce que la courbe de décision est linéaire par morceaux, il est possible de construire un ensemble de référence pour lequel la règle du plus proche voisin (PPV) est identique à la règle originale appliquée à l'ensemble de référence original. Lorsque cet ensemble est construit, l'affectation, suivant notre règle, peut tirer parti des algorithmes de classement performants pour la règle du PPV (Shamos (1975)).

Soit z l'extrémité de m segments de la courbe de décision. Généralement m est égal à deux; lorsque les enveloppes convexes ne sont pas disjointes ou lorsque le nombre de populations est supérieur à deux, m sera égal à trois pour quelques extrémités de segments (voir figures 3-5 et 3-7). Les m segments définissent les m demi-droites D_1, D_2, \dots, D_m issues de z . Nous voulons construire un ensemble fini de points z_1, \dots, z_m équidistants de z , tel que chaque secteur $D_i - D_{i-1}$ contienne un point z_i et que chaque élément de ce secteur soit plus proche (au sens de la distance usuelle) de z_i que tout autre z_j . Chaque point z_i sera incorporé à l'ensemble de référence, équivalent pour le PPV, et associé au groupe du secteur auquel il appartient.

Les points z_1, \dots, z_m étant équidistants de z , ils appartiennent à un cercle de rayon R centré en z . Tout point z_1 est dès lors entièrement déterminé par l'angle a_1 formé par les demi-droites D_1 et z, z_1 . D'autre part, les angles formés par les secteurs sont connus (b_i). Il s'en suit que les angles inconnus a_i doivent satisfaire (voir figure 4-1) :

$$b_1 - a_1 = a_2,$$

$$b_2 - a_2 = a_3,$$

...

$$b_m - a_m = a_1.$$

D'autre part :

$$b_1 + b_2 + \dots + b_m = 2\pi.$$

Supposons m impair. Substituant dans la dernière équation la valeur de a_m , puis celle de a_{m-1} , et ainsi de suite, on obtient :

$$a_1 = \pi + b_{m-1} - b_{m-3} - \dots - b_2.$$

Les autres inconnues sont traitées de la même façon. Pour obtenir une solution, il faut donc que :

$$b_2 + \dots + b_{m-1} \leq \pi.$$

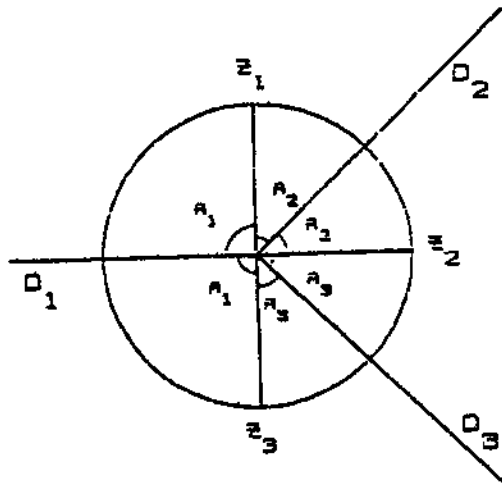


Figure 4-1. Ensemble de référence équivalent.

Si tel n'est pas le cas, nous divisons fictivement certains secteurs en deux, de telle sorte que la condition soit vérifiée. Dans ce cas, deux points seront associés, pour z , aux groupes dont le secteur a été divisé. Lorsque m est pair, un calcul semblable conduit à :

$$a_1 = b_m - b_{m-1} + \dots + b_2 - b_1 + a_1 .$$

Le système n'admet donc une solution que dans le cas très particulier où :

$$b_m - b_{m-1} + \dots + b_2 - b_1 = 0 .$$

C'est pourquoi nous diviserons un secteur en deux, lorsque m est pair, afin d'obtenir une situation où m est impair. Ainsi, lorsqu'une extrémité appartient à deux segments (le cas le plus courant), nous lui associerons trois individus étiquetés de l'ensemble de référence équivalent pour le PPV. La taille de cet ensemble de référence sera alors trois fois le nombre de segments formant la courbe de décision. Ainsi, pour un échantillon de taille 1000, l'ensemble de référence équivalent pour le PPV comprendra environ 100 individus !

5 - REFERENCES

- [1] Andrews, H.C., *Mathematical Techniques in Pattern Recognition*, Wiley, 1972.
- [2] Chow, C.K., *On Optimum Recognition Error and Reject Tradeoff*, IEEE-IT 16 : 41-46, 1970.
- [3] Degtyar, Y.U., et Finkelshteyn, M.Y., *Classification Algorithms Based on Construction of Convex Hulls of Sets*, Engineering Cybernetics 12 : 150-154, 1974.
- [4] Duda, R.O., et Hart, P.E., *Pattern Recognition and Scene Analysis*, Wiley, 1973.
- [5] Fisher, R.A., *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics 7 : 179-188, 1936.
- [6] Habbema, J.D.F., Hermans, J., et Van der Burgt, A.T., *Cases of Doubt in Allocation Problems*, Biometrika 61 : 313-324, 1974.
- [7] Hand, D.J., *Discrimination and Classification*, Wiley, 1981.
- [8] Hardy, A., et Rassin, J.P., *Une Nouvelle Approche des Problèmes de Classification Automatique*, Statistique et Analyse des Données 7 : 41-56, 1982.

- [9] Kendall, M.G., Discrimination and Classification, dans Krisnaiah, ed., Proc. Int. Symp. on Multivariate Analysis, pp. 155-185, Academic Press, 1969.
- [10] Lachenbruch, P.A., et Goldstein, M., Discriminant Analysis, Technometrics 35 : 59-85, 1979.
- [11] Lasserre, J.B., An Analytical Expression and an Algorithm for the Volume of a Convex Polyhedron in \mathbb{R}^n , à paraître dans Journal of Optimization Theory and Applications, 1984.
- [12] Levrat, P., Utilisation du Polytope Convexe d'un Nombre Fini de Points dans les Problèmes de Classement, dans Proc. 2ème Congrès AFCET-IRIA, pp. 210-216, AFCET, 1979.
- [13] Muller, D.E., et Preparata, F.P., Finding the Intersection of Two Convex Polyhedra, Theoretical Computer Science 7 : 217-236, 1978.
- [14] Rassin, J.P., Estimation de Formes Convexes du Plan, Statistique et Analyse des Données 1 : 31-46, 1979.
- [15] Ruspini, E., A New Approach to Clustering, Information and Control 15 : 22-32, 1979.
- [16] Shamos, M.I., Geometric Complexity, dans Proc. 7th ACM Conf. on the Theory of Computing, pp. 224-233, ACM, 1975.
- [17] Tomáček, I., A Generalization of the K-NN Rule, IEEE-Sys., Man., Cybern. 5 : 121-125, 1975.
- [18] Toussaint, G.T., Pattern Recognition and Geometrical Complexity, dans Proc. Fifth Int. Conf. Pattern Recognition, pp. 1334-1347, IEEE, 1980.
- [19] Wilcox, L.D., et Wilson, D.L., A Single Nearest Neighbor Rule with a Reject Option, dans Proc. Fifth Int. Conf. Pattern Recognition, pp. 98-102, IEEE, 1980.

Nous remercions A.D. Gordon (St-Andrews, Scotland) et A. Hardy (Namur, Belgique) pour les commentaires et les remarques qu'ils nous ont adressées à propos de ce travail.