# Méthodes de discrimination non paramétrique asymptotiquement efficaces au sens de Bayes

Gilles Celeux, Yves Lechevallier

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P 105 78150 Le Chesnay
France
Tél. 954 90 20

Rapports de Recherche

N° 52

# MÉTHODES DE DISCRIMINATION NON PARAMÉTRIQUE ASYMPTOTIQUEMENT EFFICACES AU SENS DE BAYES

Gilles CELEUX
Yves LECHEVALLIER

Décembre 1980

# METHODES DE DISCRIMINATION NON PARAMETRIQUE ASYMPTOTIQUEMENT EFFICACES AU SENS DE BAYES.

Gilles CELEUX - Yves LECHEVALLIER

RESUME : On présente des méthodes de discrimination qui utilisent la distance de Kolmogorov-Smirnov entre distribution de probabilité.

Les méthodes proposées se situent dans le cadre de l'approche Bayesienne de la discrimination et nous montrons qu'elles sont asymptotiquement efficaces au sens de Bayes.

D'autre part, ces méthodes, qui conduisent à la construction d'arbres de décision, sont intéressantes par leur simplicité et leur rapidité.

En particulier, on résoud le problème de la discrimination multi-classe à l'aide d'un seul arbre de décision.

On propose également une méthode utile lorsque les classes à reconnaitre sont assez mélangées.

ABSTRACT : We present discrimination methods using the Kolmogorov-Smirnov distance between probability distributions.

These methods are asymptotically Bayes' risk efficient. On another hand, they give decision trees and are interessing by their simplicity and quickness.

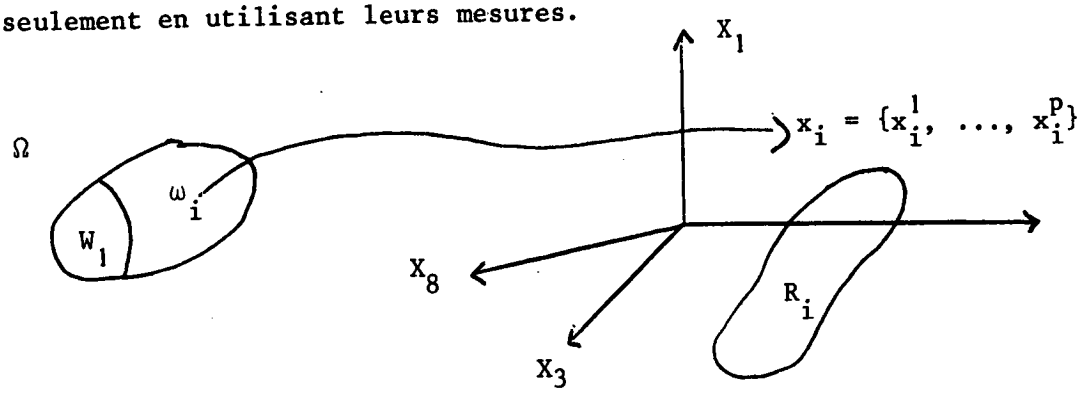
In particular, we solve the multiclass discrimination problem by using only one decision tree. We also propose a method whose aim is to separate patterns which are somewhat mixed.

PLAN :

## 1. - INTRODUCTION

Le problème de la discrimination est de classer un ensemble d'individus caractérisés par un vecteur de mesures dans plusieurs classes définies a priori.

Dans ce problème, ces individus ne peuvent pas être affectés directement mais seulement en utilisant leurs mesures.




On note :

$\{\omega_1, \ldots, \omega_N\}$ un échantillon de $\Omega$ ou chaque individu $\omega_i$ est considéré comme une observation de la population $\Omega$. $W_1, \ldots, W_K$ l'ensemble des K classes a priori.

$x_i = \{x_i^1, \ldots, x_i^p\}$ le vecteur de dimension p qui représente l'individu $\omega_i$ de $\Omega$ dans l'espace des variables. Ainsi à chaque $\omega_i$ est associé le vecteur suivant :

$$\omega_i \rightarrow (\Theta_i, x_i^1, \ldots, x_i^p)$$

où $\Theta_i = \Theta(\omega_i)$ est la classe a priori de $\omega_i$ et $x_i^j = X_j(\omega_i)$ est la valeur de $\omega_i$ sur la variable $X_j$.

Le problème de la discrimination est de construire dans l'espace des variables K régions $R_1, \ldots, R_K$ qui doivent vérifier au mieux les conditions suivantes :

1. beaucoup d'individus de $W_i$ ont leurs vecteurs de mesures dans la région $R_i$.
2. peu d'individus de $\Omega - W_i$ ont leurs vecteurs de mesures dans la région $R_i$.

Ainsi ayant K régions $R_1, \ldots, R_K$ et un échantillon $\{\omega_1, \ldots, \omega_N\}$ de $\Omega$, nous pouvons construire le tableau de contingence suivant entre les régions $R_1, \ldots, R_K$ et les classes a priori $W_1 \ldots W_K$ de $\Omega$.

| | $R_1$ | $R_2$ | | $R_K$ |
|---|---|---|---|---|
| $W_1$ | $n_{11}$ | $n_{12}$ | | $n_{1K}$ |
| $W_2$ | $n_{21}$ | $n_{22}$ | | $n_{2K}$ |
| | | | | |
| $W_K$ | $n_{K1}$ | $n_{K2}$ | | $n_{KK}$ |

où $n_{ij}$ est le nombre d'individus de l'échantillon $\{\omega_1, \ldots, \omega_N\}$ appartenant à la classe a priori $W_i$ et classés dans la région $R_j$.

Le pourcentage de mauvaise classification $\frac{n_{ij}}{N}$ de la classe a priori $W_i$ et de la région $R_j$ est une estimation de la probabilité de mauvaise classification entre $W_i$ et $R_j$.

$$P(W_i \cap R_j) = P(R_j/W_i) \times P(W_i)$$

où $P(W_i)$ est la probabilité a priori de $W_i$ et $P(R_j/W_i)$ représente la probabilité conditionnelle de $R_j$ sachant $W_i$.

Le calcul du côut de mauvaise classification quand les régions $R_1, \ldots, R_K$ sont sélectionnés se fait ainsi : On note $C(W_i, R_j)$ le coût de mauvaise classification quand la décision de classer un individu dans la $j^{ième}$ classe est prise alors qu'il appartient à la $i^{ème}$ classe. Alors le coût moyen de mauvaise classification (risque de Bayes) est :

$$\sum_{\substack{i=1,K \\ j=1,K}} C(W_i, R_j)\, P(W_i \cap R_j)$$

Les méthodes de discrimination faisant décroitre ce risque sont appelées méthodes bayesiennes et sont étudiées ici.

## 2. - APPROCHE BAYESIENNE DE LA DISCRIMINATION

Pour simplifier nous considérons deux classes $W_1$ et $W_2$ avec les probabilités a priori $\pi_1 = P(W_1)$ et $\pi_2 = P(W_2)$. Nous supposons que les densités de probabilité, par rapport à la mesure de Lebesgue dans l'espace des variables, des classes $W_1$ et $W_2$ existent.

On veut construire la fonction de décision qui sépare l'espace des variables en deux régions $R_1$ et $R_2$. On note C (1,2) le coût de classer une observation de $W_1$ dans $R_2$ et C (2,1) le coût de classer une observation de $W_2$ dans $R_1$.

Ceci nous permet de construire le tableau des coûts suivants :

| | $R_1$ | $R_2$ |
|---|---|---|
| $W_1$ | 0 | C(1,2) |
| $W_2$ | C(2,1) | 0 |

On a la table des probabilités correspondante :

| | $R_1$ | $R_2$ |
|---|---|---|
| $W_1$ | P(1,1) | P(1,2) |
| $W_2$ | P(2,1) | P(2,2) |

Ainsi le coût moyen de mauvaise classification s'écrit :

$P(2,1).C(2,1) + P(1,2).C(1,2)$.

La connaissance des probabilités a priori $\pi_i$ de $W_i$ et des fonctions de densité $f_i$ de $W_i$ dans l'espace des variables permet de calculer ce coût.

La probabilité de bien classer un individu sachant qu'il est issu de $W_1$ est $\int_{R_1} f_1(x)\,dx$ avec $dx = dx_1, \ldots dx_p$.

La probabilité de mal classer cet individu est $\int_{R_2} f_1(x)\,dx$

d'où

$p(1,2) = \pi_1 \int_{R_2} f_1(x)\,dx$.

Le coût moyen de mauvaise classification est alors :

$$\pi_1 \cdot C(1,2) \cdot \int_{R_2} f_1(x)\,dx + \pi_2 \cdot C(2,1) \cdot \int_{R_1} f_2(x)\,dx,$$

Une procédure qui minimise ce coût est une procédure bayésienne.

Lorsque les probabilités a priori sont connues la règle de décision consiste à minimiser (si on pose $\mathbb{R}^p = R_1 \cup R_2$) :

$$\pi_1 \cdot C(1,2) \cdot \int_{R^p - R_1} f_1(x)\,dx + \pi_2 \cdot C(2,1) \cdot \int_{R_1} f_2(x)\,dx$$

soit $\pi_1\, C(1,2) \cdot \int_{R^p} f_1(x)\,dx + \int_{R_1} [\pi_2 \cdot C(2,1)\, f_2(x) - \pi_1\, C(1,2)\, f_1(x)]\,dx$

Le premier terme étant constant la règle de décision est de choisir $R_1$ tel que :

$$R_1 = \{x \in R^p / \pi_1 \cdot C(1,2)\, f_1(x) > \pi_2 \cdot C(2,1)\, f_2(x)\}$$

soit

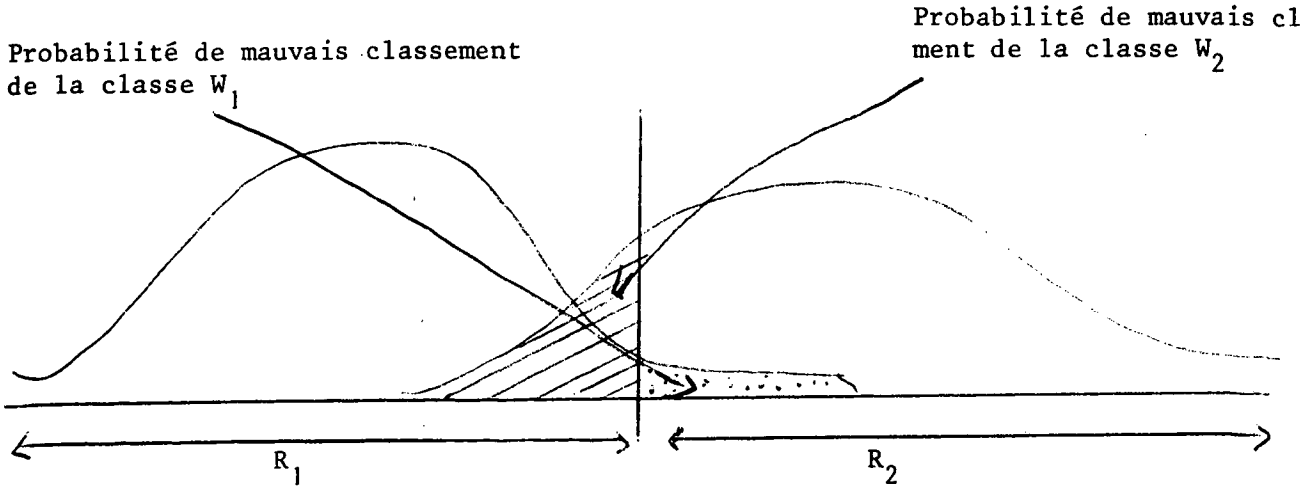
$$R_1 = \{x \in R^p / \frac{f_1(x)}{f_2(x)} \geq \frac{C(2,1)}{C(1,2)} \frac{\pi_2}{\pi_1}\}$$

et

$$R_2 = \{x \in R^p / \frac{f_1(x)}{f_2(x)} < \frac{C(2,1)}{C(1,2)} \frac{\pi_2}{\pi_1}\}$$

Dans le cas où $\pi_1\, C(1,2) = \pi_2\, C(2,1)$ il vient :

$$R_1 = \{x \in R^p / f_1(x) \geq f_2(x)\}$$



Les fonctions de densité étant souvent inconnues partiellement ou totalement le problème nécessite d'estimer ces fonctions de densité,

Trois voies sont utilisées :

1) On suppose que les fonctions de densité de $W_1$ et de $W_2$ sont des distributions normales de vecteurs moyennes $\mu_1$ et $\mu_2$ et de matrice variance $\Sigma$.

Cette approche a été utilisée par Wald (1944). [Wal 44].

2) La procédure des k plus proches voisins permet d'estimer les fonctions de densité

Cette approche a été utilisée en premier par Fix et Hodges (1951) [FiH 51].

3) On peut utiliser d'autres méthodes non paramétriques pour estimer les fonctions de densité. (cf. Henrichon et Fu (1969) [HeF 69], Friedman (1977)[Fri 77]

2.1. - Analyse discriminante linéaire (approche classique)

Dans le cas de deux populations multinormales de même matrice variance $\Sigma$ et de moyennes $\mu_1$ et $\mu_2$ la fonction de densité de la $i^{ème}$ classe est :

$$f_i(x) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)' \Sigma^{-1} (x-\mu_i)\right]$$

d'où

$$\log \frac{f_1(x)}{f_2(x)} = -\frac{1}{2}\left[(x-\mu_1)' \Sigma^{-1} (x-\mu_1) - (x-\mu_2)' \Sigma^{-1} (x-\mu_2)\right]$$

$$= x' \Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2) \Sigma^{-1} (\mu_1-\mu_2)$$

Cette fonction étant linéaire en x, les régions sont données par :

$$R_1 = \{x \in R^p / x' \Sigma^{-1} (\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2) \Sigma^{-1} (\mu_1-\mu_2) \geq \text{Log } k\}$$

et

$$R_2 = \{x \in R^p / x' \Sigma^{-1} (\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2) \Sigma^{-1}(\mu_1-\mu_2) < \text{Log } k\}$$

avec k défini par :

$$k = \frac{\pi_2 \, C(2,1)}{\pi_1 \, C(1,2)}$$

Dans beaucoup d'applications les paramètres des fonctions de densité sont inconnues. Dans ce cas, nous disposons d'un échantillon de chaque classe normale et nous utilisons cette information pour estimer ces paramètres.

Supposons que nous ayons un échantillon $x_1^{(1)}, \ldots, x_{n_1}^{(1)}$ de $W_1$ et $x_1^{(2)}, \ldots, x_{n_2}^{(2)}$ de $W_2$.

Dans ce cas les estimateurs du maximum de vraisemblance des paramètres $\mu_i$ sont :

$$\bar{x}^{(i)} = \sum_{\alpha=1}^{n_i} \frac{x_\alpha^{(i)}}{n_i}$$

et celui de $\Sigma$ est S qui vérifie

$$(n_1 + n_2 - 2)S = \sum_{i=1}^{2} \sum_{\alpha=1}^{n_i} (x_\alpha^{(i)} - \bar{x}^{(i)})' (x_\alpha^{(i)} - \bar{x}^{(i)})$$

En substituant ces estimations dans l'expression de la fonction de décision, on obtient :

$$R_1 = \{x \in R^P / x' S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) - \frac{1}{2} [(\bar{x}^{(1)} + \bar{x}^{(2)}) S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)})] \geq \text{Log } k\}$$

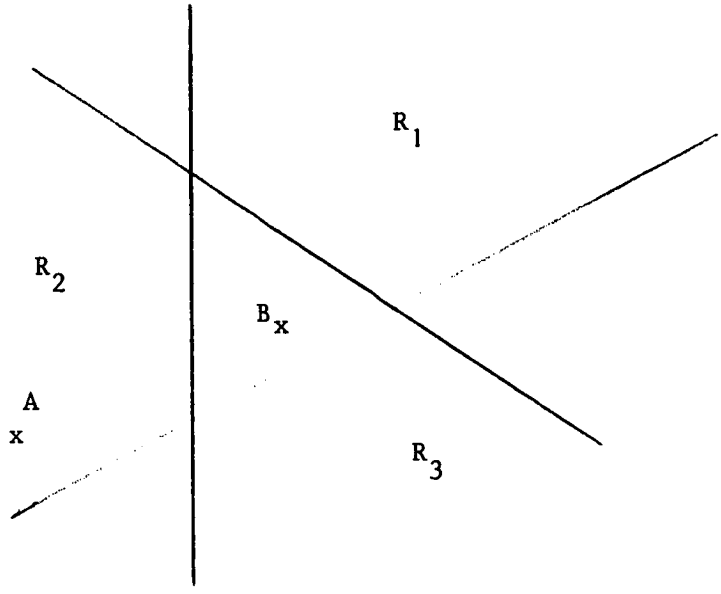
Le premier terme de cette fonction discriminante a été suggéré par Fisher en 1936.

Comme la fonction discriminante f est linéaire en x, ($f(x) = ax + b$) la surface de séparation entre $R_1$ et $R_2$ est un hyperplan d'équation

$f(x) = ax + b = 0$.

Une nouvelle observation x est assignée à la classe $W_1$ si $f(x) \geq 0$ et à la classe $W_2$ si $f(x) < 0$. Pour chaque paire de classes un hyperplan peut être construit. Pour K classes, il y a $\frac{1}{2} K(K-1)$ hyperplans séparateurs.

Exemple : K = 3




Dans cet exemple, le point A est classé dans la région $R_2$, mais le point B ne peut être classé car il doit être classé dans la classe 2 si on compare $W_1$ et $W_2$, dans la classe 1 si on compare $W_1$ et $W_3$ et dans la classe 3 si on compare $W_2$ et $W_3$. Ainsi par cette règle, ce point est impossible à classer.

Remarque :

Dans cette approche, on suppose que les classes sont des populations normales. Cette hypothèse assez forte n'est pas toujours acceptable et demande à être vérifiée au préalable.

## 2.2. - Règle de décision des k plus proches voisins

Il est clair que les résultats de la règle de décision dépendent de l'estimation des fonctions de densité. Si elles ont plusieurs modes le problème de l'estimation des fonctions de densité est difficile. Une des approches consiste à supposer les densités composées de mélanges de lois normales. [Par 62].

Une autre approche est d'estimer les densités en chaque point que l'on désire classer, c'est l'approche des k plus proches voisins. Cette technique suppose que la distance d entre deux individus de $R^P$ soit connue. Ayant un échantillon $\{x_1, \ldots, x_n\}$, la fonction de densité $f_i$ est estimée au point x de la manière suivante :

Une suite de régions $\{B_k\}$ k = 1,n centrées sur x est construite. Chaque région a un volume $\emptyset_k$ et contient k points de l'échantillon.

Parmi ces k objets, $k_i$ sont issus de $W_i$.

On estime $f_i(x)$ par $\hat{f}_i(x) = \frac{k_i}{n_i \emptyset_k}$ , k étant fixé , avec $n_i$ = card $W_i$

A partir de ces estimations, on prend la décision de classer x dans la classe $W_j$ si :

$$\frac{\hat{f}_j(x)}{\hat{f}_i(x)} \geq \frac{c(i,j)\pi_i}{c(j,i)\pi_j} \qquad \forall\ i \neq j.$$

Remarques :

La procédure des k plus proches voisins utilise une distance définie sur $R^P$ et les résultats dépendent du choix de cette distance.

De plus, la recherche des k plus proches voisins d'un point x dans un espace de dimension P est souvent très couteuse.

## 3. - APPROCHE NON PARAMETRIQUE DANS LE CAS DE 2 FAMILLES

Dans cet article nous allons étudié une procédure paramétrique dû à J. FRIEDMAN [Fri 77] intéressante par sa simplicité et sa rapidité. Diverses extensions sont proposées.

### 3.1. Principes de la méthode dans le cas de 2 familles et d'une seule variable

Soit W la population totale, $W_1$ et $W_2$ les deux familles avec card $W_1$ = n, card $W_2$ = m = N - n. Dans le cas, où une seule variable décrit la population, on pose $f_1(x)$ et $f_2(x)$ les densités continues de probabilité des 2 familles et $F_1(x)$ et $F_2(x)$ les fonctions de répartition correspondantes.

On notera $\pi_1$ (resp. $\pi_2$) la probabilité a priori pour un élément d'appartenir à la famille $W_1$ (resp. $W_2$). On a $\pi_1 + \pi_2 = 1$. On notera $\ell_1$ (resp. $\ell_2$) le coût de mauvaise classification d'un élément de $W_1$ (resp. $W_2$) dans $W_2$ (resp. $W_1$).

Soit $z_{N+1}$ un individu supplémentaire, on veut classer $z_{N+1}$ dans $W_1$ ou $W_2$ à partir de l'échantillon $z_1, \ldots, z_N$, en utilisant le point coupure c de la manière suivante :

si $z_{N+1} \leq c$, il est affecté dans la population "inférieure", $W_1$ par exemple,
si $z_{N+1} > c$, il est affecté dans la population "supérieure", soit $W_2$.

Dans ce cadre, on veut choisir c de telle manière à minimiser le risque de Bayes R de mauvaise classification de cet individu test. On se restreindra ici au cas où $\pi_1 \ell_1 = \pi_2 \ell_2$.

La valeur du risque de Bayes R(z) de mal classer $z_{N+1}$ en utilisant un point quelconque z pour effectuer la coupure est :

$$R(z) = \ell_1 \pi_1 (1 - F_1(z)) + \ell_2 \pi_2 F_2(z).$$

On en déduit que sous l'hypothèse $\pi_1 \ell_1 = \pi_2 \ell_2$, on a l'équivalence :

$$\min_z R(z) \iff \max_z (F_1(z) - F_2(z)).$$

Comme on ne connaît pas a priori la population inférieure et la population supérieure, le point qui minimise le risque de Bayes est le point c tel que :

$$D(c) = \sup_z D(z) = \sup_z |F_1(z) - F_2(z)|.$$

La quantité D(c) n'est autre que la distance de Kolmogorov-Smirnov entre les deux distributions et est une mesure bien connue de la séparabilité des 2 fonctions de répartitions.

En pratique, $F_1$ et $F_2$ ne sont pas connues et on les estime par les fonctions de répartition empiriques définies ainsi :

$$\hat{F}_1(x) = \begin{cases} 0 \text{ si } x < x_1^1 \\ \frac{k}{n} \text{ si } x_k^1 \leq x < x_{k+1}^1 \\ 1 \text{ si } x_n^1 \leq x \end{cases} \qquad \hat{F}_2(x) = \begin{cases} 0 \text{ si } x < x_1^2 \\ \frac{k}{m} \text{ si } x_k^2 \leq x < x_{k+1}^2 \\ 1 \text{ si } x_m^2 \leq x \end{cases}$$

formules dans lesquelles $x_k^i$ est le $k^{\text{ième}}$ point de la famille $W_i$ ($i = 1, 2$) les points étant rangés par ordre croissant.

On estime $D(c)$ par $\hat{D}(\hat{c}) = \sup_z |\hat{F}_1(z) - \hat{F}_2(z)|$

Le risque effectif de mauvaise classification résultant de cette procédure est $R(\hat{c})$. Le risque "estimé" de mauvaise classification est

$$\hat{R}(\hat{c}) = \ell_1 \pi_1 (1 - \hat{F}_1(\hat{c})) + \ell_2 \pi_2 \hat{F}_2(\hat{c}))$$

Sous la restriction $\pi_1 \ell_1 = \pi_2 \ell_2$, la proposition 1 revient à montrer que le risque "estimé" $\hat{R}(\hat{c})$ converge en probabilité vers le risque de Bayes ($R(c)$).

Le proposition 2 (cf. STO 54) revient à montrer que le risque effectif $R(\hat{c})$ converge en probabilité vers le risque de Bayes.

3.1.1. Proposition 1

$\hat{D}(\hat{c})$ converge en probabilité vers $D(c)$.

Démonstration : Dire que $\hat{D}(\hat{c})$ converge vers $D(c)$ en probabilité signifie que :

$\forall \varepsilon > 0 \quad \forall \eta > 0 \; \exists N$ tq pour tout $n > N$, $P[|\hat{D}(\hat{c}) - D(c)| > \varepsilon] < \eta$

Ceci est équivalent à la conjonction des deux propositions suivantes :

(p.1.1) $\forall \varepsilon > 0 \; \forall \eta > 0 \; \exists N_1$ tq pour tout $n > N_1, P[(D(c) - \hat{D}(\hat{c})) > \varepsilon] < \eta$

(p.1.2) $\forall \varepsilon > 0 \; \forall \eta > 0 \; \exists N_2$ tq pour tout $n > N_2, P[(\hat{D}(\hat{c}) - D(c)) > \varepsilon] < \eta$

Démonstration de p.1.1 :

Pour tout z, $\hat{D}(z) = |\hat{F}_1(z) - \hat{F}_2(z)|$ converge en probabilité vers D(z) car la fonction de répartition empirique converge en probabilité vers le fonction de répartition.

En particulier, $\hat{D}(c)$ converge en probabilité vers D(c). On peut alors écrire :
$\forall \varepsilon > 0$, $\forall n > 0$ $\exists N_1$ $t_q$ pour tout $n > N_1$ . $P[\hat{D}(c) < D(c) - \varepsilon] < \eta$

Or il est clair que $\hat{D}(\hat{c}) = \max_z |\hat{F}_1(z) - \hat{F}_2(z)| > |\hat{F}_1(c) - \hat{F}_2(c)| = \hat{D}(c)$,
on a donc bien a fortiori :

$$P[\hat{D}(\hat{c}) < D(c) - \varepsilon] < \eta$$

Démonstration de p 1.2 :

On a :

$$\hat{D}(\hat{c}) - D(c) = \max_z |\hat{F}_1(z) - \hat{F}_2(z)| - \max_z |F_1(z) - F_2(z)|$$

$$\hat{D}(\hat{c}) - D(c) \leq \max_z [|\hat{F}_1(z) - \hat{F}_2(z)| - |F_1(z) - F_2(z)|]$$

et on peut écrire :

$$|\hat{F}_1(z) - \hat{F}_2(z)| - |F_1(z) - F_2(z)| = |\hat{F}_1(z) - \hat{F}_2(z) - F_1(z) + F_1(z)| - |F_1(z) - F_2(z)|$$

d'où :

$$|\hat{F}_1(z) - \hat{F}_2(z)| - |F_1(z) - F_2(z)| \leq |\hat{F}_1(z) - F_1(z)| + |F_1(z) - \hat{F}_2(z)| - |F_1(z) - F_2(z)|$$

$$= |\hat{F}_1(z) - F_1(z)| + |F_1(z) - \hat{F}_2(z) + F_2(z) - F_2(z)| - |F_1(z) - F_2(z)|$$

$$\leq |\hat{F}_1(z) - F_1(z)| + |F_2(z) - \hat{F}_2(z)| + |F_1(z) - F_2(z)| - |F_1(z) - F_2(z)|.$$

Donc :

$$|\hat{F}_1(z) - \hat{F}_2(z)| - |F_1(z) - F_2(z)| \leq |\hat{F}_1(z) - F_1(z)| + |\hat{F}_2(z) - F_2(z)|$$

On en déduit :

$$\hat{D}(\hat{c}) - D(c) \leq \max_z [|\hat{F}_1(z) - F_1(z)| + |\hat{F}_2(z) - F_2(z)|]$$

$$\Rightarrow \hat{D}(\hat{c}) - D(c) \leq \max_z [|\hat{F}_1(z) - F_1(z)|] + \max_z [|\hat{F}_2(z) - F_2(z)|]$$

et d'après le théorème de Glyvenko-Cantelli, pour i = 1, 2 $\max_z |\hat{F}_i(z) - F_i(z)|$ tend vers 0 presque sûrement donc a fortiori en probabilité. Donc : pour $\varepsilon > 0$ et $\eta > 0$ fixés, $\exists$ $N_2$ tel que pour tout $n > N_2$ $P[\hat{D}(\hat{c}) - D(c)) > \varepsilon] < \eta$.

Finalement

$\forall \varepsilon > 0$, $\forall \eta > 0$ $\exists$ $N_3 = \sup(N_1, N_2)$ tq pour tout $n > N_3$ $P[|\hat{D}(\hat{c}) - D(c)| > \varepsilon] < \eta$ et donc $\hat{D}(\hat{c})$ converge en probabilité vers D(c).

3.1.2. Proposition 2

$D(\hat{c})$ converge en probabilité vers D(c).

Démonstration :

On a :

$$|D(\hat{c}) - D(c)| = |D(\hat{c}) - \hat{D}(\hat{c}) + \hat{D}(\hat{c}) - D(c)| \leq |\hat{D}(\hat{c}) - D(c)| + |\hat{D}(\hat{c}) - D(\hat{c})|$$

La proposition précédente permet d'affirmer que $|\hat{D}(\hat{c}) - D(c)|$ converge en probabilité vers 0.

Examinons maintenant le deuxième terme : $|\hat{D}(\hat{c}) - D(\hat{c})|$.

On sait que $\forall z$, $\hat{D}(z)$ converge en probabilité vers $D(z)$. En particulier $\hat{D}(\hat{c})$ convergence en probabilité vers $D(\hat{c})$ d'où :

$|\hat{D}(\hat{c}) - D(\hat{c})|$ tend vers 0 en probabilité et finalement l'inégalité $|D(\hat{c}) - D(c)| \leq |\hat{D}(\hat{c}) - D(c)| + |\hat{D}(\hat{c}) - D(\hat{c})|$ assure la convergence en probabilité de $D(\hat{c})$ vers $D(c)$.

## 3.2. - Extension de la procédure à plusieurs variables

Une extension naturelle de la procédure précédente au cas multivariable est d'effectuer la coupure pour la variable où la distance de Kolmogorov-Smirnov entre les deux distributions est la plus grande. On réapplique ensuite la procédure à chaque sous-population jusqu'à ce qu'on rencontre un test d'arrêt.

Plus précisément, l'algorithme peut se résumer ainsi :

On calcule pour toutes les variables la quantité :

$D(c_j) = \max_z |\hat{F}_1^j(z) - \hat{F}_2^j(z)|$ ($F_i^j$ (resp. $F_2^j$) représentant la fonction de répartition de la famille $W_1$ (resp. $W_2$) pour la variable $j$ ) et l'on effectue la coupure pour la variable $j_*$ tel que :

$$D(c_{j*}) = \max_j (D(c_j))$$

La coupure se fait au point $c_{j*}$.

Si l'un des deux sous-échantillons satisfait au test d'arrêt, il est affecté à l'une des deux classes (celle qui est majoritaire dans le sous échantillon) et on obtient ainsi une cellule terminale. Sinon on reprend la procédure à partir de ce sous-échantillon. En particulier, on recalcule $D(c_j)$ même pour la variable $j$ choisie au(x) pas précédent(s), en effet dans le cas où $f_1(x)$ et/ou $f_2(x)$ sont multimodes, il se peut qu'une seule coupure ne fournisse pas une bonne discrimination.

3.2.1. Le test d'arrêt

Il reste à définir le test d'arrêt. L'assignement à une des familles est fait sur la base de l'estimation du rapport $f_1/f_2$. Le cardinal de chaque famille dans les sous-échantillons doit être assez grand pour permettre une estimation raisonnable du rapport des densités. Aussi le partitionnement s'arrêtera chaque fois que le partitionnement suivant n'assurera pas des échantillons de taille minimum pour chaque famille.

Le choix du nombre k minimum doit être déterminé par l'utilisateur et dépend du problème posé. k doit croître avec n, mais plus lentement que n. En effet, Gordon et Ohlsen montrent [Go078] que la procédure décrite ici est efficace asymptotiquement au sens de Bayes si

$$\lim_n \frac{k(n)}{n} = 0 \text{ et } \lim_n \frac{k(n)}{\sqrt{n}} = +\infty.$$

Autrement dit, toujours sous la restriction $\ell_1\pi_1 = \ell_2\pi_2$, lorsque la taille des deux échantillons devient grande, le risque de Bayes de classement d'un nouvel individu approche, avec une probabilité arbitrairement proche de 1, le risque de Bayes basé sur une procédure de Bayes construite à partir d'une connaissance complète des distributions $F_1$ et $F_2$. Il s'agit en fait de la généralisation au cas multivariables de la proposition 2.

3.2.2. Affectation de nouveaux individus

Le partitionnement par cette procédure conduit à un arbre de décision binaire. Un sous-échantillon à chaque étape est représenté par un noeud de l'arbre. Le sommet de l'arbre redonne l'échantillon tout entier. Les deux successeurs de chaque noeud non terminal représentent les 2 sous-ensembles définis par partitionnement. Les noeuds terminaux représentent les cellules terminales.

La règle de classification d'un nouvel individu est donc simple. On l'assigne à la classe caractérisant la cellule dans laquelle il tombe. Partant du sommet, l'individu descend l'arbre jusqu'à ce qu'il arrive à un noeud terminal. Si celui-ci représente une classe unique, il est affecté à cette classe. Si le noeud représente un mélange de classes, il est assigné à la classe majoritaire. En descendant

l'arbre, la décision d'aller à gauche ou à droite est prise ainsi : si $x_{j*} \leq c_{j*}$, on va au successeur gauche, sinon on va au successeur de droite. Ici $j_*$ et $c_{j*}$ représente la variable et la coupure correspondante à ce noeud.

### 3.2.3. Remarques

La procédure garantit le meilleur partitionnement à chaque noeud de l'arbre. Mais elle ne passe pas en revue tous les choix de séquences de coupures possibles. Ainsi le partitionnement de l'espace obtenu est sous optimal au sens statistique. Ceci étant, sauf cas particulier, il ne semble pas que cette restriction altère les performances de l'algorithme et de toute façon, l'examen de toutes les suites de coupures n'est pas réalisable même pour de petits échantillons.

Dans ce qui précède, nous avons supposé que $\pi_1 \ell_1 = \pi_2 \ell_2$, ce qui garantit pour la procédure la propriété d'efficacité assymptotique au sens de Bayes. Ceci étant, le critère choisi, ici la distance de Kolmogorov-Smirnov entre les 2 distributions, se justifie en dehors des considérations de convergence, par son bon pouvoir de séparation entre 2 distributions.

Si l'on désire conserver la propriété de convergence du risque de Bayes et si l'hypothèse $\pi_1 \ell_1 = \pi_2 \ell_2$ n'apparaît pas satisfaisante, on peut définir le critère à partir d'une estimation de $\pi_1, \ell_1, \pi_2, \ell_2$.

On doit alors à chaque pas de l'algorithme trouver c qui minimise le risque de Bayes $R(z) = \ell_1 \pi_1 (1 - F_1(z)) + \ell_2 \pi_2 F_2(z)$ et en pratique on doit considérer $\hat{R}(z) = \hat{\ell}_1 \hat{\pi}_1 (1 - F_1(z)) + \hat{\ell}_2 \hat{\pi}_2 F_2(z)$ où $\hat{\pi}_1, \hat{\ell}_2, \hat{\pi}_1, \hat{\ell}_2$ représentent les estimations des valeurs théoriques. La modification introduite dans l'algorithme garantit l'efficacité assymptotique au sens du risque de Bayes [GoO 78].

## 4. - APPROCHE NON PARAMETRIQUE DANS LE CAS MULTI-CLASSES

### 4.1. Introduction

Une extension possible de cette procédure à des problèmes à plus de 2 classes consiste à les considérer comme une succession de problèmes à 2 classes.

Si on a K classes (K>2) à discriminer, on construit les K arbres de décision correspondant à chaque classe (contre toutes les autres).

Pour classer un individu, Friedman préconise de procéder ainsi : [Fri 77].

L'individu à classer tombe dans K cellules terminales. Pour chacune de ces cellules, on calcule $C_j$ = nombre d'éléments de la classe j dans la cellule de l'arbre j et $O_j$ = nombre d'éléments n'appartenant pas à la classe j dans cette même cellule.

On affecte l'individu dans la classe j qui maximise $C_j - O_j$ sur les cellules.

Cette procédure peut s'avérer peu efficace dans le cas ou $\forall$ j = 1, ..., K $C_j - O_j < 0$. Il y a alors un risque d'affecter l'individu dans une classe à laquelle il a peu de chances d'appartenir.

Aussi nous proposons dans ce cadre la règle d'affectation suivante [Cel 78].

L'individu à classer tombe dans K cellule terminales. Pour chacune de ces cellules, on calcule $C_j - O_j$.

On affecte l'individu à la classe j* qui maximise $C_j - O_j$, <u>à condition</u> que $C_{j*} - O_{j*} \geq 0$.

Sinon, on calcule pour les K cellules la quantité $M_j - N_j$ avec $M_j$ = nombre d'éléments de la classe qui a le plus grand nombre d'éléments dans la cellule de l'arbre j.

$N_j$ = nombre d'éléments n'appartenant pas à cette classe dans cette même cellule.

On affecte, alors, l'individu à la classe j qui maximise $M_j - N_j$ sur les K cellules terminales.

Dans les différentes applications que nous avons faites de cet algorithme, cette règle d'affectation s'avère systèmatiquement meilleure que la précédente.

Ceci étant, la stratégie qui consiste à ramener un problème multi-classe à une succession de problèmes à deux classes ne nous paraît pas la meilleure.

D'une part, la procédure d'affectation de nouveaux individus perd en rapidité.

D'autre part, les décisions offertes par cette méthode peuvent être contradictoires si l'on considère chaque arbre de décision pris séparément :

Soit, par exemple, trois classes A, B, C. Au vue de l'arbre de décision de la classe A, un individu peut être affecté à la classe A et au vu de l'arbre de la classe B, être affecté à la classe B, etc...

Bien sur, les règles d'affectation que nous avons indiquées permettent de trancher, mais ces procédures ne nous paraissent pas très fiables dans de tels cas douteux.

Aussi la méthode suivante nous semble préférable.

## 4.2. - Approche multiclasse avec un seul arbre de décision

### 4.2.1. Introduction

Le problème est de construire un seul arbre de décision associé à un ensemble de K classes. Nous noterons $W_1, \ldots, W_K$ les K familles et $\Pi_1, \ldots, \Pi_K$ les probabilités a priori associés à ces familles et $F_1, \ldots, F_K$ leurs fonctions de répartition théoriques. Soit $W = \{W_1, \ldots, W_K\}$ alors pour $A \in P(W)$ la fonction de répartition théorique est :

$$F_A(x) = \frac{1}{\Pi_A} \sum_{W_i \in A} \Pi_i \cdot F_i(x) \quad \text{avec} \quad \Pi_A = \sum_{W_i \in A} \Pi_i .$$ Ainsi pour un ensemble

A de familles a priori le risque de Bayes au point z est égal à :

$R(z) = \ell_A \pi_A (1 - F_A(z)) + \ell_{\bar{A}} \pi_{\bar{A}} F_{\bar{A}}(z)$ avec $\bar{A} = W - A$ d'où $\pi_{\bar{A}} = \sum_{W_i \in A} \pi_i$ et si

on suppose $\ell_A = \ell_{\bar{A}}$ on a $R(z) = \pi_A + \pi_{\bar{A}} F_{\bar{A}}(z) - \pi_A F_A(z)$.

$\pi_A$ est une constante et comme on ne connait pas à priori la population inférieure et la population supérieure le point c qui minimise le risque de Bayes est :

$$D(c) = \sup_z \left| \pi_A F_A(z) - \pi_{\bar{A}} F_{\bar{A}}(z) \right| .$$

On estime $D(c)$ par $\hat{D}(c) = \sup_Z \left| \pi_A \hat{F}_A(z) - \pi_{\bar{A}} \hat{F}_{\bar{A}}(z) \right|$

car on suppose que les probabilités $\pi_A$ et $\pi_{\bar{A}}$ sont connues. On note $\mathcal{A}$ l'ensemble des partitions en deux classes de $\{W_1, \ldots, W_k^{\mathcal{A}}\}$.

4.2.2. Proposition

$D(\hat{c})$ converge en probabilité vers $D(c)$ ce qui revient à démontrer que le risque effectif $R(\hat{c})$ converge vers le risque $R(c)$.

Démonstration

$$\left| D(\hat{c}) - D(c) \right| = \left| \pi_A F_A(\hat{c}) - \pi_{\bar{A}} F_{\bar{A}}(\hat{c}) - \pi_A F_A(c) - \pi_{\bar{A}} F_{\bar{A}}(c) \right|$$

$$\left| D(\hat{c}) - D(c) \right| \leq \left| \pi_A F_A(\hat{c}) - \pi_A F_A(c) \right| + \left| \pi_{\bar{A}} F_{\bar{A}}(\hat{c}) - \pi_{\bar{A}} F_{\bar{A}}(c) \right|$$

$$\left| D(\hat{c}) - D(c) \right| \leq \sum_{W_i \in A} \pi_i \left| F_i(\hat{c}) - F_i(c) \right| + \sum_{W_i \notin A} \pi_i \left| F_i(\hat{c}) - F_i(c) \right|$$

$F_i(\hat{c})$ converge en probabilité vers $F_i(c)$ car $\hat{F}_i(\hat{c})$ converge en probabilité vers $F_i(\hat{c})$ et $\hat{F}_i(\hat{c})$ converge presque surement vers $F_i(c)$ ce d'après les propositions 1 et 2 du paragraphe 3.1.1.

4.2.3. Extension de la procédure à plusieurs variables

Cette extension est identique à celle du paragraphe 3.2.. Ainsi on calcule pour toutes les variables la quantité :

$$R(c_j) = \max_{A \in \mathcal{A}} \max_{x} \left| \pi_{\bar{A}} F_{\bar{A}}(x) - \pi_A F_A(x) \right|$$

et l'on effectue la coupure pour la variable $j_*$ tel que :

$$R(c_{j*}) = \max_{j} \{R(c_j)\}.$$

4.2.4. Considération numérique

La recherche de la classe A réunion d'un ensemble de classes a priori ne nécessite pas l'énumération complète de tous les cas possibles d'un regroupement en deux classes d'un ensemble de k classes a priori. Cette recherche ne nécessite que la construction de K-1 regroupements car il suffit d'ordonner les classes en fonction de la valeur de leur fonction de répartition en un point donné. La solution optimale est nécessairement dans les (K-1) regroupements formés par les (K-1) coupures possibles : $c_1, \ldots, c_{K-1}$.

Ceci peut se démontrer comme ceci : soit K valeurs ordonnées $x_1 < \ldots < x_K$ et on suppose que la partition en deux classes de ces K valeurs ne forme pas deux intervalles disjoints. Dans ce cas il existe une valeur $x_\ell$ n'appartenant pas à la première classe (classe des petites valeurs) et inférieure à une valeur $x_{\ell'}$ appartenant à cette première classe. En calculant les moyennes des deux classes :

$$\bar{x}_1 = \frac{1}{n} \sum_{i \in c_1} x_i \quad \text{et} \quad \bar{x}_2 = \frac{1}{n-k} \sum_{i \in c_2} x_i$$

nous obtenons un écart $\Delta = \bar{x}_2 - \bar{x}_1$. En permutant les deux valeurs $x_\ell$ et $x_{\ell'}$ on obtient deux nouvelles classes avec les moyennes suivantes :

$$\bar{x}'_1 = \frac{1}{n} \sum_{i \in c'_1} x_i \text{ avec } c'_1 = c_1 - \{x_{\ell'}\} + \{x_\ell\}$$

et comme $x_{\ell'} > x_\ell$ on a :

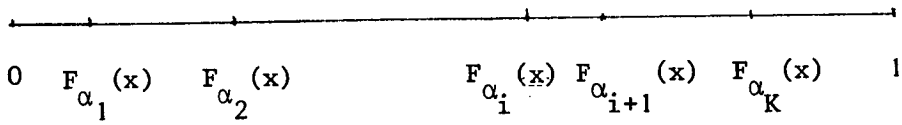
$$\bar{x}'_1 < \bar{x}_1$$

$$\bar{x}'_2 = \frac{1}{n-k} \sum_{i \in c'_2} x_i \text{ avec } c'_2 = c_2 - \{x_\ell\} + \{x_{\ell'}\}$$

et comme $x_\ell < x_{\ell'}$, on a :

$$\bar{x}'_2 > \bar{x}_2 \quad \text{d'où}$$

$$|\bar{x}'_2 - \bar{x}'_1| > |\bar{x}_2 - \bar{x}_1|$$




$F_{\alpha_1}(x) < c_1 < F_{\alpha_2}(x)$ d'où la classe $A = \{W_{\alpha_1}\}$

$F_{\alpha_2}(x) < c_2 < F_{\alpha_2}(x)$ d'ou la classe $A = \{W_{\alpha_1}, W_{\alpha_2}\}$

$\vdots$

$F_{\alpha_i}(x) < c_i < F_{\alpha_{i+1}}(x)$ d'où la classe $A = \{W_{\alpha_1}, \ldots, W_{\alpha_i}\}$

$\vdots$

$F_{\alpha_{K-1}}(x) < c_{K-1} < F_{\alpha_K}(x)$ d'où la classe $A = \{W_{\alpha_1}, \ldots, W_{\alpha_{K-1}}\}$

Ainsi la recherche de la solution optimale est de complexité linéaire en fonction de K.

### 4.2.5. Exemples d'applications

L'ensemble des données choisies pour cette application est l'ensemble des données de Fisher [Fis 35] concernant 3 populations d'Iris. Chaque population d'Iris est échantillonnée avec 50 individus. Dans Fisher [Fis 35] est utilisé l'analyse discriminante linéaire sur les variables. Il construit une fonction linéaire discriminante entre deux populations d'Iris ; cette fonction linéaire n'est pas nécessaire pour séparer la population des Iris Setosa des autres Iris car la moyenne de ces Iris sur la variable n° 4 (largeur de la pétale) est de 0,246 et l'étendue est de 0,2 à 0,6 alors que pour les autres Iris le minimum est de 1,0. Ceci démontre que cette 4ème variable est assez discriminante en elle-même pour séparer les Iris Setosa des autres. Par contre, la discrimination entre les Iris Versicolor et Virginica ne peut se faire parfaitement par une seule variable.

La construction d'une fonction discriminante linéaire dépendant de plusieurs variables permet une très bonne séparation entre ces deux populations. Cependant l'utilisation de nos arbres de décision binaires permet d'obtenir une aussi bonne séparation.

Cet ensemble de données a été utilisé par Kendall [Ken 66] avec sa "distribution-free method". De la même façon que notre méthode, cette méthode utilise uniquement le caractère ordonné d'une variable quantitative. Cette méthode découpe la variable en trois zones ; les deux zones extrêmes définissent l'intervalle de décision d'affection à une population : la zone centrale est l'intervalle d'indécision. Sur l'échantillon test les zones extrêmes ne doivent contenir que des individus de la même population. Par exemple les variables n° 3 (longueur de la pétale) et n° 4 (largeur de la pétale) sont des variables relativement discriminantes pour les deux populations versicolor (Vers) et Virginica (Virg) (voir Table 1 ci-dessous).

| Longueur de la pétale (PL) | | | Largeur de la pétale (PW) | | |
|---|---|---|---|---|---|
| Valeur | Vers | Virg | Valeur | Vers | Virg |
| ≤ 4.3 | 25 | | ≤ 1.3 | 28 | |
| 4.4 | 4 | | 1.4 | 7 | 1 |
| 4.5 | 7 | 1 | 1.5 | 10 | 2 |
| 4.6 | 3 | | 1.6 | 3 | 1 |
| 4.7 | 5 | | 1.7 | 1 | 1 |
| 4.8 | 2 | 2 | 1.8 | 1 | 11 |
| 4.9 | 2 | 3 | ≥ 1.9 | | 34 |
| 5.0 | 1 | 3 | | | |
| 5.1 | 1 | 7 | | | |
| ≥ 5.2 | | 34 | | | |
| Total | 50 | 50 | Total | 50 | 50 |

d'où la règle de décision suivante :

PL ≤ 4.4. affectation à la population Versicolor
PL ≥ 5.2. affectation à la population Virginica
4.5. ≤ PL ≤ 5.1. indécision (choisir une nouvelle variable).

Ainsi il reste 37 cas d'indécision. Dans ce cas la variable PW est maintenant choisie et nous obtenons la règle de décision suivante :

avec 4.5. ≤ PL ≤ 5.1.
PW ≤ 1.4. affectation à la population Versicolor
alors si PW ≥ 1.9. affectation à la population Virginica
1.5. ≤ PW ≤ 1.9. indécision et choix d'une nouvelle variable.

Maintenant il reste 22 cas d'indécision. La variable largeur de la sépale (SW) est maintenant choisie et nous avons une nouvelle fonction de décision :

SW ≥ 3.1. affectation Versicolor
SW < 3.1. indécision, choix d'une nouvelle variable.

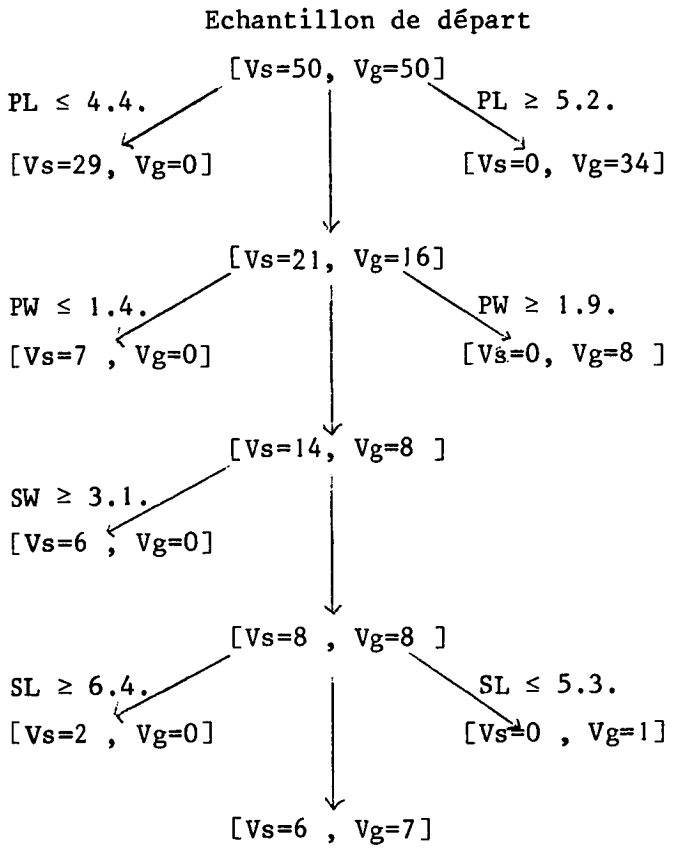
Il reste 16 cas d'indécision et la dernière variable longueur de la sépale (SL) permet de construire la fonction de décision suivante :

SL ≥ 6.4. affectation à la population Versicolor
SL ≤ 5.3. affectation à la population Virginica
5.3. ≤ SL ≤ 6.3. indécision.

Aucune amélioration n'est possible et nous avons 13 cas d'indécision.

L'arbre de décision de cette "distribution-free method" est la suivante :

Versicolor = Vs
Virginica = Vg



De même que la méthode précédente, notre méthode du § 4.2 a comme but la construction d'un arbre de décision mais la coupure d'une variable s'effectue en fonction du risque de Bayes. Ainsi notre méthode ne permet pas d'obtenir directement des zones ne contenant que d'individus de la même population. Dans l'exemple de Fisher les deux populations ont la même probabilité a priori (p=0.50) et on peut dire que ce coût d'erreur est le même pour chaque population. Ainsi minimiser le risque de Bayes revient à maximiser le test de Smirnoff-Kolmogorov. Ce test de Smirnoff-Kolmogorov est calculé pour toutes les coupures possibles sur toutes les variables. Ce calcul est repris ci-dessous pour les variables n° 3 et n° 4, ces variables étant les plus discriminantes.

| Longueur de la pétale (PL) | | | | | | Largeur de la pétale (PW) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Valeur | Vers | fré-quence | Virg | fré-quence | test SK | Valeur | Vers | fré-quence | Virg | fré-quence | test SK |
| ≤ 4.3. | 25 | 0,50 | | 0,00 | 0,50 | ≤ 1.3 | 28 | 0,56 | | 0,00 | 0,56 |
| 4.4 | 4 | 0,58 | | 0,00 | 0,58 | 1.4 | 7 | 0,70 | 1 | 0,02 | 0,68 |
| 4.5 | 7 | 0,72 | 1 | 0,02 | 0,70 | 1.5 | 10 | 0,90 | 2 | 0,06 | 0,86 |
| 4.6 | 3 | 0,78 | | | 0,76 | 1.6 | 3 | 0,96 | 1 | 0,08 | 0,88 |
| 4.7 | 5 | 0,88 | | | 0,86 | 1.7 | 1 | 0,98 | 1 | 0,10 | 0,88 |
| 4.8 | 2 | 0,92 | 2 | 0n06 | 0,86 | 1.8 | 1 | 1,00 | 11 | 0,32 | 0,68 |
| 4.9 | 2 | 0,96 | 3 | 0,12 | 0,84 | ≥ 1.9 | | 1,00 | 34 | 1,00 | 0 |
| 5.0 | 1 | 0,98 | 3 | 0,18 | 0,80 | | | | | | |
| 5.1 | 1 | 1,00 | 7 | 0,32 | 0,68 | | | | | | |
| ≥ 5.2 | | 1,00 | 34 | 1,00 | 0,00 | | | | | | |
| Total | 50 | | 50 | | | Total | 50 | | 50 | | |

d'où pour la variable PL

$\max_x |\hat{F}_1(x) - \hat{F}_2(x)| = 0,86$ et la valeur à la convergence est 4,7, d'où pour la variable PW

$\max_x |\hat{F}_1(x) - \hat{F}_2(x)| = 0,88$ et la valeur à la convergence est 1,6 ou 1,7.

La variable discriminante est PW et la coupure est la valeur 1,6.

| | $V_S$ | $Y_g$ |
|---|---|---|
| PW ≤ 1,6 | 48 | 4 |
| PW > 1,6 | 2 | 46 |

A la deuxième étape la variable la plus discriminante de l'ensemble des Iris vérifiant PW ≤ 1,6 est la variable PL(longueur de la pétale) avec un test de Smirnoff-Kolmogorov de 0,979 et la coupure en 4.9. La fonction de décision est la suivante :

PW ≤ 1,6 [Vs = 48, Vg = 4]

PL ≤ 4.9 [Vs = 47, Vg = 0]

PL > 4.9 [Vs = 1, Vg = 4]

L'ensemble défini par PW ≤ 1,6 et PL ≤ 4.9 est un ensemble terminal car il y a une discrimination parfaite des Versicolors. Par contre, il faut décomposer l'ensemble PW ≤ 1,6 et PL > 4.9 .

La variable la plus discriminante de l'ensemble des Iris vérifiant PW > 1,6 est aussi la variable PL avec un test égal à 0,826 et la coupure en 5. La fonction de décision est la suivante :

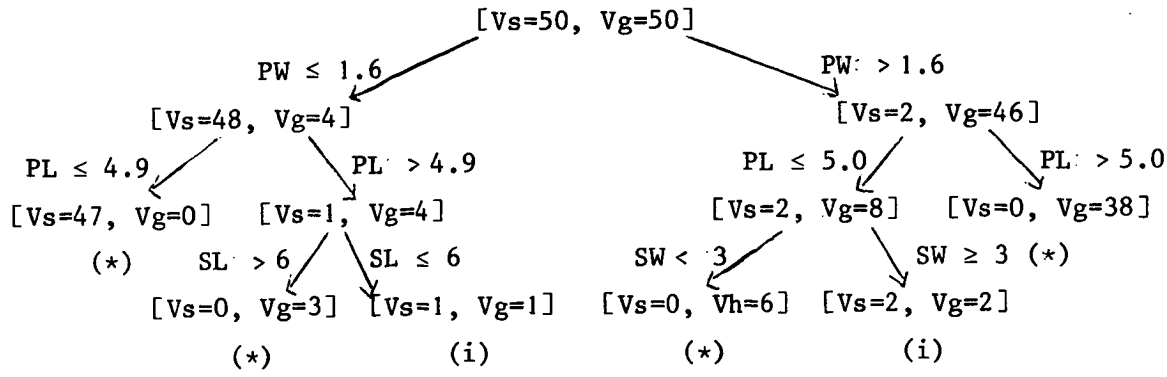
PW > 1,6 [Vs = 2, Vg = 46]

PL ≤ 5.0 [Vs = 2, Vg = 8]

PL > 5.0 [Vs = 0, Vg = 38]

Ici l'ensemble défini par PW > 1,6 et PL > 5.0 est un ensemble terminal et caractérise les Iris Virginica.

A la troisième étape nous reprennons les deux ensembles non discriminants. Avec l'ensemble [PW ≤ 1,6 et PL > 4.9] la variable la plus discriminante est la longueur de la sépale (SL) et la valeur de la coupure est 6. Avec l'ensemble [PW > 1,6 et PL ≤ 5.0] la variable la plus discriminante est la largeur de la sépale (SW) et la valeur de la coupure est 3.0, l'arbre de décision de cette méthode est le suivant :




Les ensembles marqués d'une étoile (*) sont des ensembles composés uniquement d'individus appartenant à la même population.

Avec cette méthode il reste deux ensembles d'indécision (i), ces deux ensembles comprennent 6 individus. Sur cet ensemble de données cette méthode s'avère meilleure que la méthode proposée par Kendall [Ken 66].

La méthode de Kendall construit à chaque étape de la segmentation des ensembles totalement discriminés.

Dans notre méthode, ces ensembles totalement discriminés n'apparaissent qu'à la deuxième étape car notre méthode recherche des ensembles assez discriminants au sens du risque de Bayes.

Le risque de Bayes semble être un critère bien adapté dans la recherche de groupes discriminants.

## 5. - UNE VARIANTE : ARBRES DE DECISION TERNAIRES

L'algorithme que nous avons présenté vaut par sa rapidité de mise en oeuvre et la facilité d'interprétation des résultats.

La principale critique que l'on peut lui faire est son manque de finesse. S'il est très efficace et agréable pour reconnaître des classes relativement bien séparées, il est moins performant pour reconnaitre des classes "proches" l'une de l'autre.

Dans cette partie, nous présentons une variante dont le but est de fournir une discrimination plus fine tout en conservant la facilité d'interprétation des résultats de la méthode précédente et qui reste performante au point de vue de la rapidité.

### 5.1. - Principes de la méthode dans le cas de 2 familles

L'idée est la suivante : au lieu de construire des arbres de décision binaires, on va construire des arbres de décision à 3 branches, la branche du milieu étant une branche d'indécision.

Le formalisme de cette variante est tout à fait analogue à celui de la méthode de base. Aussi nous l'exposons en suivant le même plan que précédemment.

#### 5.1.1. Cas d'une seule variable ( p = 1)

Les définitions sont les mêmes qu'au début du § 3. Mais en plus des coûts $\ell_1$ (resp. $\ell_2$) de mauvaise classification pour un élément de $W_1$ (resp. $W_2$), on introduit un coût de non décision $\ell'_1$ (resp. $\ell'_2$) pour un élément de $W_1$ (resp. $W_2$).

Le problème s'énonce alors ainsi :

Au vu de l'échantillon $z_1, \ldots, z_n$, on veut classer un individu supplémentaire $z_{n+1}$ dans $W_1$ ou $W_2$ en utilisant deux points coupure $c_1$ et $c_2$ ($c_1 \leq c_2$) de la manière suivante :

si $z_{n+1} \leq c_1$, il est affecté dans la population "inférieure", $W_1$ par exemple.

si $z_{n+1} > c_2$, il est affecté dans la population "supérieure", soit $W_2$.

si $c_1 < z_{n+1} \leq c_2$, on ne prend pas décision, ce qui signifie que la variable considérée ne permet pas de décider de l'affectation de $z_{n+1}$.

Dans ce cadre, on veut choisir $c_1$ et $c_2$ de manière à minimiser le risque de Bayes $R(z_1, z_2)$ de mal classer un individu en utilisant les deux points coupures $z_1$ et $z_2$ $(z_1 \leq z_2)$.

Là encore, on se restreindra au cas $\pi_1 \ell_1 = \pi_2 \ell_2$. Et de manière cohérente avec cette dernière restriction, on supposera de plus que $\pi_1 \ell'_1 = \pi_2 \ell'_2$.

Enfin, on posera $\ell_1 = a\ell'_1$ avec $a \geq 1$, ce qui entraîne que $\ell_2 = a\ell'_2$ car de $\pi_1 \ell_1 = \pi_2 \ell_2$ on tire $\pi_1 a \ell'_1 = \pi_2 \ell_2 = a \pi_2 \ell'_2$ puisque $\pi_1 \ell'_1 = \pi_2 \ell'_2$.

Notons que prendre $a \geq 1$ est naturel : le coût de mal classer un élément de $W_1$ est plus grand que de ne pas classer cet individu à l'aide de la variable considérée.

Le risque de Bayes s'écrit :

$$R(z_1, z_2) = \pi_1 \ell_1 (1 - F_1(z_2)) + \pi_2 \ell_2 F_2(z_1) + \pi_1 \ell'_1 (F_1(z_2) - F_1(z_1)) +$$

$$+ \pi_2 \ell'_2 (F_2(z_2) - F_2(z_1))$$

Dans le cas où $\pi_1 \ell_1 = \pi_2 \ell_2 = a \pi_1 \ell'_1 = a \pi_2 \ell'_2$, il vient :

$$R(z_1,z_2) = \pi_1 \ell'_1 [a + F_1(z_2)(1-a) - F_1(z_1) + F_2(z_1)(a-1) + F_2(z_2)].$$

On a alors l'équivalence :

$$\min_{\substack{z_1,z_2 \\ z_1 \leq z_2}} R(z_1,z_2) \Longleftrightarrow \max_{\substack{z_1,z_2 \\ z_1 \leq z_2}} [F_1(z_2)(a-1) + F_1(z_1) - F_2(z_1)(a-1) - F_2(z_2)]$$

En fait, on ne connait pas a priori la population inférieure. Mais nous verrons que l'algorithme détermine au début la population inférieure de la manière suivante :

$W_1$ (resp. $W_2$) est la population inférieure si

$$D(c) = \sup_z |F_1(z) - F_2(z)| = F_1(c) - F_2(c) \text{ (resp. } D(c) = F_2(c) - F_1(c))$$

Dans toute la suite on supposera donc que $W_1$ est la population inférieure.

La variante consiste donc à chercher les deux points coupures $c_1$ et $c_2$ qui maximise le critère :

$$D(z_1,z_2) = [F_1(z_2)(a-1) + F_1(z_1) - F_2(z_1)(a-1) - F_2(z_2)]$$

En pratique, on maximisera

$$\hat{D}(z_1,z_2) = [\hat{F}_1(z_2)(a-1) + \hat{F}_1(z_1) - \hat{F}_2(z_1)(a-1) - \hat{F}_2(z_2)]$$

et donc on estime $D(c_1,c_2) = \sup_{\substack{z_1,z_2 \\ z_1 \le z_2}} D(z_1,z_2)$ par $\hat{D}(c_1,c_2) = \sup_{\substack{z_1,z_2 \\ z_1 \le z_2}} \hat{D}(z_1,z_2)$

De manière analogue au cas binaire, on a les propriétés de convergence suivantes :

Proposition : $\hat{D}(\hat{c}_1,\hat{c}_2)$ converge vers $D(c_1,c_2)$ en probabilité.

Proposition : $D(\hat{c}_1,\hat{c}_2)$ converge en probabilité vers $D(c_1,c_2)$.

On en déduit le résultat suivant :

Le risque effectif de mauvaise classification $R(\hat{c}_1,\hat{c}_2)$ converge en probabilité vers le risque de Bayes $R(c_1,c_2)$.

Démonstration de la **première** proposition :

$\forall (z_1,z_2)\ \hat{D}(z_1,z_2)$ converge en probabilité vers $D(z_1,z_2)$ car la fonction de répartition empirique converge en probabilité vers la fonction de répartition.

En particulier $\hat{D}(c_1,c_2)$ converge en probabilité vers $D(c_1,c_2)$. On peut alors écrire :

$\forall \varepsilon > 0 \quad \forall \eta > 0\ \exists N$, tq pour tout $n > N_1$

$$P[\hat{D}(c_1,c_2) < D(c_1,c_2) - \varepsilon] < \eta$$

Or par construction, $\hat{D}(\hat{c}_1, \hat{c}_2) > \hat{D}(c_1,c_2)$ d'où $P[\hat{D}(\hat{c}_1, \hat{c}_2) < D(c_1,c_2) - \varepsilon] < \eta$ pour tout $n > N_1$. D'autre part :

$$\hat{D}(\hat{c}_1,\hat{c}_2) - D(c_1,c_2) = \sup_{\substack{z_1,z_2 \\ z_1 \le z_2}} [(a-1)\,\hat{F}_1(z_2) + \hat{F}_1(z_1) - (a-1)\,\hat{F}_2(z_1) - \hat{F}_2(z_2)]$$

$$- \sup_{\substack{z_1,z_2 \\ z_1 \le z_2}} [(a-1)\,F_1(z_2) + F_1(z_1) - F_2(z_1)(a-1) - F_2(z_2)]$$

$$\hat{D}(\hat{c}_1,\hat{c}_2) - D(c_1,c_2) \le \sup_{\substack{z_1,z_2 \\ z_1 \le z_2}} \{[(a-1)\,\hat{F}_1(z_2) + \hat{F}_1(z_1) - (a-1)\,\hat{F}_2(z_1) - \hat{F}_2(z_2)] -$$

$$- [(a-1)\,F_1(z_2) + F_1(z_1) - (a-1)\,F_2(z_1) - F_1(z_2)]\}$$

On a :

$$[(a-1)\,\hat{F}_1(z_2) + \hat{F}_1(z_1) - (a-1)\,\hat{F}_2(z_1) - \hat{F}_2(z_2)] - [(a-1)F_1(z_2) + F_1(z_1) - (a-1)F_2(z_1) -$$

$$- F_2(z_2)] = [(a-1)(\hat{F}_1(z_2) - F_1(z_2)) + (\hat{F}_1(z_1) - F_1(z_1)) - (a-1)(\hat{F}_2(z_1) -$$

$$- F_2(z_1)) - (\hat{F}_2(z_2) - F_2(z_2))] \le (a-1)\,|\hat{F}_1(z_2) - F_1(z_2)| +$$

$$+ |\hat{F}_1(z_1) - F_1(z_1)| + (a-1)\,|\hat{F}_2(z_1) - F_2(z_1)| + |\hat{F}_2(z_2) - F_2(z_2)|.$$

On en déduit :

$$\hat{D}(\hat{c}_1,\hat{c}_2) - D(c_1,c_2) \leq \sup_{z_2} (a-1)\ |\hat{F}_1(z_2) - F_1(z_2)| + \sup_{z_1} |\hat{F}_1(z_1) - F_1(z_1)| +$$

$$+ \sup_{z_1} (a-1)\ |\hat{F}_2(z_1) - F_2(z_1)| + \sup_{z_2} |\hat{F}_2(z_2) - F_2(z_2)|.$$

et d'après le théorème de Glyvenko-Cantelli, on a donc :

$\forall \varepsilon > 0, \forall \eta > 0, \exists N_2$ tq pour tout $n > N_2$

$$P\left[(\hat{D}(\hat{c}_1,\hat{c}_2) - D(c_1,c_2)) > \varepsilon\right] < \eta$$

Finalement :

$\forall \varepsilon > 0, \forall \eta > 0 \quad \exists N_3 = \sup(N_1, N_2)$ tq pour tout $n > N_3$

$$P\left[|\hat{D}(\hat{c}_1,\hat{c}_2) - D(c_1,c_2)| > \varepsilon\right] < \eta .$$

Démonstration de la deuxième proposition :

On a :

$$|D(\hat{c}_1,\hat{c}_2) - D(c_1,c_2)| \leq |\hat{D}(\hat{c}_1,\hat{c}_2) - D(\hat{c}_1,\hat{c}_2)|$$

$$+ |\hat{D}(\hat{c}_1,\hat{c}_2) - D(c_1,c_2)|.$$

D'après la proposition précédente $\hat{D}(\hat{c}_1,\hat{c}_2)$ converge en probabilité vers $D(c_1,c_2)$ et $\hat{D}(\hat{c}_1,\hat{c}_2)$ converge en probabilité vers $D(\hat{c}_1,\hat{c}_2)$ puisque la fonction de répartition empirique converge en probabilité vers la fonction de répartition. On en déduit aisément le résultat annoncé.

5.1.2. Extension de la méthode à plusieurs variables

L'algorithme est alors le même que dans le cas binaire.

A chaque pas on sélectionne la variable qui maximise le critère :

$$\hat{D}(\hat{c}_1,\hat{c}_2) = \sup_{\substack{z_1,z_2 \\ z_1 \leq z_2}} [(a-1)\,\hat{F}_1(z_2) + \hat{F}_1(z_1) - (a-1)\,\hat{F}_2(z_1) - \hat{F}_2(z_2)]$$

et on effectue les deux coupures aux points $\hat{c}_1$ et $\hat{c}_2$.

On réitère la procédure sur chacun des sous-échantillons ainsi obtenus. Le test d'arrêt est le même que dans le cas binaire.

Le partitionnement par cette procédure conduit à un arbre de décision où de chaque sommet part deux ou trois branches selon les cas. Il part deux branches si au sommet considéré les points coupures $\hat{c}_1$ et $\hat{c}_2$ sont égaux, trois branches si $\hat{c}_1 < \hat{c}_2$.

Ce dernier cas se produit lorsque la variable j sélectionnée au sommet considéré ne permet pas de discriminer entre les deux familles tous les individus dont la coordonnée $x_j$ pour cette variable vérifie $\hat{c}_1 < x_j \leq \hat{c}_2$.

L'introduction de cette branche du "milieu" permet ainsi d'éviter une affectation peu sûre d'individus à l'une des 2 familles définies a priori.

Lorsque la branche du milieu disparaît, c'est-à-dire lorsque pour le sommet considéré et la variable sélectionnée correspondante on a $\hat{c}_1 = \hat{c}_2$, le critère s'écrit :

$$\hat{D}(\hat{c}_1,\hat{c}_1) = a \sup_z |\hat{F}_1(z) - \hat{F}_2(z)|$$

L'unique point coupure est alors le même point que l'on obtient dans le cas binaire.

Ainsi, si les familles a priori sont bien séparées, cette variante donnera les mêmes résultats que l'algorithme de base : de chaque sommet de l'arbre de décision, il ne partira que deux branches.

5.2. - Cas multi-classes : (L > 2)

Il est clair que quel que soit la manière d'envisager le problème multi-classe les algorithmes que nous avons présenté se généralise sans peine pour cette variante.

5.3. - Le choix de a :

Dans la variante proposée, le nombre $a \geq 1$ tel que $\ell_1 = a\,\ell'_1$ est un paramètre de l'algorithme qui doit être défini par l'utilisateur.

Remarquons tout d'abord que le critère est une fonction croissante de a.

Cas limites :

Si a = 1, (coût de mauvaise classification égal au coût de non décision), le critère s'écrit :

$$\hat{D}(\hat{c}_1,\hat{c}_2) = \sup_{\substack{z_1,z_2 \\ z_1 \leq z_2}} [\hat{F}_1(z_1) - \hat{F}_2(z_1)]$$

d'où

$$\hat{D}(\hat{c}_1,\hat{c}_2) = \sup_{z} |\hat{F}_1(z) - \hat{F}_2(z)|$$

On retombe alors sur la procédure utilisant la distance de Kolmogorov-Smirnov entre les deux distributions et donc sur un arbre de décision binaire.

Si $a \to \infty$ (coût de non décision négligeable devant le coût de mauvaise classification).

On a alors $\ell'_1 \to 0$ et $\ell'_2 \to 0$ car $\ell_1$ est fixé. Le critère devient :

$$\hat{D}(\hat{c}_1,\hat{c}_2) = \sup_{\substack{z_1,z_2 \\ z_1 \leq z_2}} [\hat{F}_1(z_2) - \hat{F}_1(z_1)]$$

Il suffit alors de prendre :

$$\hat{c}_1 \leq x^{(1)}$$

et

$$\hat{c}_2 > x^{(n)}$$

$x^{(1)}$ étant la plus petite valeur rencontré sur l'échantillon,
$x^{(n)}$ étant la plus grande valeur rencontré sur l'échantillon.

On a alors :

$$D(\hat{c}_1,\hat{c}_2) = 1$$

Tous les individus vérifient $\hat{c}_1 < x_i < \hat{c}_2$ et sont classés dans la classe du milieu. Ce qui conduit à ne jamais prendre de décision.

Entre ces deux extrêmes, l'utilisateur doit choisir a de manière à réaliser un compromis entre la qualité de la discrimination d'une part et la simplicité des résultats.

Si a est trop grand, l'arbre de décision aura de nombreux sommets et sera difficile à interpréter.

Si a est trop petit, l'arbre sera peu différent d'un arbre de décision binaise et l'on risque alors d'obtenir des résultats moins bons.

D'autre part, on a le résultat suivant :

Proposition : Pour $1 \leq a \leq 2$, on a :

$$\hat{D}(\hat{c}_1,\hat{c}_2) = a \sup_z |\hat{F}_1(z) - \hat{F}_2(z)| = a\, D(c)$$

Démonstration :

$\forall (z_1,z_2)$ avec $z_1 \leq z_2$,

$$D(z_1,z_2) = (a-1)\,\hat{F}_1(z_2) + \hat{F}_1(z_1) - (a-1)\,\hat{F}_2(z_1) - \hat{F}_2(z_2)$$

$$D(z_1,z_2) = (a-1)(\hat{F}_1(z_2) - \hat{F}_2(z_2)) + (a-1)(\hat{F}_1(z_1) - \hat{F}_2(z_1))$$

$$+ (a-2)(\hat{F}_2(z_2) - \hat{F}_1(z_1))$$

d'où

$$D(z_1,z_2) \leq 2(a-1)\, D(c) + a - 2$$

et

$2(a-1)D(c) + a - 2 \leq a\, D(c)$ dès que $a \leq 2$, d'où $D(z_1,z_2) \leq a\, D(c)$ si $a \leq 2$.

Ainsi l'on est amené à prendre $a > 2$ pour éviter de retomber sur le critère de Kolmogorov-Smirnov.

Nous pensons que prendre $a = 3$, soit $\ell_1 = 3\ \ell'_1$ est une assez bonne solution.

En pratique, l'utilisateur pourra faire varier a au cours de différents essais de manière à obtenir un arbre de décision qui le satisfasse en regard du problème posé.

5.4. - Intérêt de cette variante

Cette variante permet d'affiner la discrimination lorsque les familles a priori sont difficiles à séparer.

D'un point de vue théorique, cette variante généralise l'algorithme de base. En effet, lorsque les familles a priori seront bien séparées, les branches du mi-

lieu de l'arbre de décision disparaitront et l'on retombera exactement sur l'arbre de décision binaire.

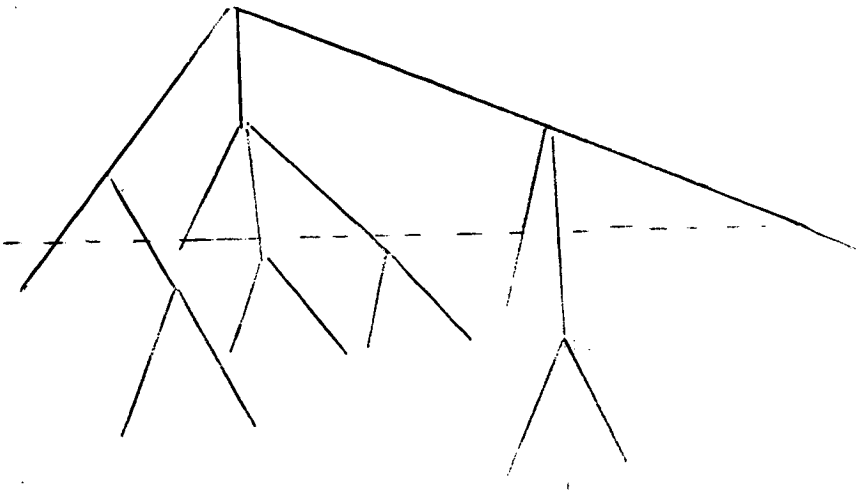
D'un point de vue pratique, dans ce dernier cas, cette variante allongera inutilement le temps calcul. Ceci étant, il est difficile de savoir a priori si les familles à reconnaitre sont bien séparées ou non.

Un autre intérêt de cette variante résulte dans le fait qu'elle facilite le dialogue homme machine.

En effet, cette méthode de reconnaissance automatique ne décide pas forcément de l'affectation de tous les individus. Cet état de fait correspond à une réalité. Dans pratiquement tous les problèmes de discrimination, il existe des individus qui sont mal classés quel que soit la qualité des fonctions de décision. Cette méthode permet de les déceler facilement. Le spécialiste (le médecin par exemple) devra envisager de les affecter à une famille par d'autres moyens.

A ce propos, nous voulons faire la remarque suivante : les derniers sommets peuvent s'avérer moins fiables car plus sujets aux fluctuations d'échantillonage que les premiers sommets.

Ainsi dans les cas où les arbres de décision obtenus comportent de multiples niveaux, il peut être intéressant de les couper à un certain niveau qui ne soit pas forcément celui des cellules terminales. Par la même, on peut améliorer la sureté de la discrimination.



Coupure de l'arbre à un niveau qui fournit 8 classes dont 2 de non décision

## 5.5. - Considérations numériques

Dans ce paragraphe, nous allons voir qu'il n'est pas nécessaire de passer en revue tous les couples $(z_1,z_2)$ pour optimiser le critère $\hat{D}(\hat{c}_1,\hat{c}_2) = \sup_{\substack{z_1,z_2 \\ z_1 \le z_2}} \hat{D}(z_1,z_2)$.

On pose card $W_1 = n$

card $W_2 = m$

On examine le critère pour une variable donnée.

On suppose toujours que :

$$\sup_z |\hat{F}_1(z) - \hat{F}_2(z)| = \hat{F}_1(c) - \hat{F}_2(c).$$

Autrement dit $W_1$ est la population inférieure et $W_2$ est la population supérieure. On suppose d'autre part que $a > 2$.

Au point c est associé le couple $(n_c,m_c)$, vérifiant $n_c,m_c \in \mathbb{N}$, $0 \le n_c \le n$, $0 \le m_c \le m$ et tel que :

$n_c = \text{card } \{\omega \in W_1 \ / \ x(\omega) < c\}$ , $m_c = \text{card } \{\omega \in W_2 \ / \ x(\omega) < c\}$

On a : $\hat{F}_1(c) - \hat{F}_2(c) = \frac{n_c}{n} - \frac{m_c}{m}$

soit

$$C(c) = \sup_z \hat{D}(z,z) = a\left(\frac{n_c}{n} - \frac{m_c}{m}\right) \text{ avec } a > 2.$$

De manière analogue, à chaque point coupure $c_1$ et $c_2$ maximisant le critère : $(\hat{D}(c_1,c_2) = \sup_{\substack{z_1,z_2 \\ z_1 \le z_2}} \hat{D}(z_1,z_2))$ est associé un couple d'entiers $(n_{c_1}, m_{c_1})$ pour $c_1$ et $(n_{c_2}, m_{c_2})$ pour $c_2$ vérifiant $0 \le n_{c_1} \le n_{c_2} \le n$ et $0 \le m_{c_1} \le m_{c_2} \le m$ et tels que

$$\hat{D}(c_1,c_2) = (a-1)\frac{n_{c_2}}{n} + \frac{n_{c_1}}{n} - (a-1)\frac{m_{c_1}}{m} - \frac{m_{c_2}}{m}$$

Proposition :

Pour a > 2, le couple $(c_1,c_2)$ vérifiant $\hat{D}(c_1,c_2) = \sup_{\substack{z_1,z_2 \\ z_1 \le z_2}} \hat{D}(z_1,z_2)$ est tel que :

$c_1 \le c \le c_2$ où c est défini par $\sup_z |\hat{F}_1(z) - \hat{F}_2(z)| = \hat{F}_1(c) - \hat{F}_2(c) = \frac{n_c}{n} - \frac{m_c}{m}$

Démonstration :

Soit $(z_1,z_2)$ tel que $z_1 < z_2 \le c$.

On a :

$$D(z_1,z_2) = (a-1)\left[\frac{n_{z_2}}{n} - \frac{m_{z_1}}{m}\right] + \left[\frac{n_{z_1}}{n} - \frac{m_{z_2}}{m}\right]$$

$$D(z_1,c) - D(z_1,z_2) = (a-1)\left[\left(\frac{n_c}{n} - \frac{n_{z_2}}{n}\right)\right] - \frac{m_c}{m} + \frac{m_{z_2}}{m}$$

$$z_2 \le c \Rightarrow \frac{n_{z_2}}{n} \le \frac{n_c}{n} \text{ et } a - 1 > 1$$

d'où

$$D(z_1,c) - D(z_1,z_2) \ge \frac{n_c}{n} - \frac{n_{z_2}}{n} - \frac{m_c}{m} + \frac{m_{z_2}}{m} = \left[\frac{n_c}{n} - \frac{m_c}{m}\right] - \left[\frac{n_{z_2}}{n} - \frac{m_{z_2}}{m}\right]$$

si $\left[\frac{n_{z_2}}{n} - \frac{m_{z_2}}{m}\right] \le 0$, on a $D(z_1,c) - D(z_1,z_2) \ge 0$

si $\left[\frac{n_{z_2}}{n} - \frac{m_{z_2}}{m}\right] > 0$, on a de toute façon, par définition de c : $\left[\frac{n_{z_2}}{n} - \frac{m_{z_2}}{m}\right] \le \left[\frac{n_c}{n} - \frac{m_c}{m}\right]$.

On en déduit que $D(z_1,c) \ge D(z_1,z_2)$ $\forall$ $z_2 \le c$.

On montrerait de manière analogue que
$D(c,z_2) \ge D(z_1,z_2)$ $\forall (z_1,z_2)$ vérifiant $c \le z_1 \le z_2$. D'où la proposition annoncée.

Proposition :

Le couple $(c_1,c_2)$ vérifiant $\hat{D}(c_1,c_2) = \sup_{\substack{z_1,z_2 \\ z_1 \geq z_2}} \hat{D}(z_1,z_2)$ est tel que :

$\hat{D}(c_1,c) = \sup_{z_1 \leq c} \hat{D}(z_1,c)$ et $\hat{D}(c,c_2) = \sup_{z_2 \geq c} \hat{D}(c,z_2)$.

Démonstration :

Soit $c_1$ et $c_2$ vérifiant $\hat{D}(c_1,c) = \sup_{z_1 \leq c} \hat{D}(z_1,c)$ et $\hat{D}(c,c_2) = \sup_{z_2 \geq c} \hat{D}(c,z_2)$

on peut poser : $n_{c_1} = n_c - n_1$, $m_{c_1} = m_c - m_1$ $\quad n_1, m_1 \in N$

$n_{c_2} = n_c + n_2$, $m_{c_2} = m_c + m_2$ $\quad n_2, m_2 \in N$

et on a $\hat{D}(c_1,c) = a(\frac{n_c}{n} - \frac{m_c}{m}) + (a-1)\frac{m_1}{m} - \frac{n_1}{n}$

$\hat{D}(c,c_2) = a(\frac{n_c}{n} - \frac{m_c}{m}) + (a-1)\frac{n_2}{n} - \frac{m_2}{m}$

Soit maintenant $(c'_1,c'_2)$ vérifiant $\hat{D}(c'_1,c'_2) = \sup_{\substack{z_1,z_2 \\ z_1 \leq z_2}} \hat{D}(z_1,z_2)$.

D'après la proposition précédente, on a :

$n_{c'_1} = n_c - n'_1$, $m_{c'_1} = m_c - m'_1$, $n_{c'_2} = n_c + n'_2$, $m_{c'_2} = m_c + m'_2$

$n'_1, m'_1, n'_2, m'_2 \in \mathbb{N}$

Et on peut écrire :

$\hat{D}(c'_1,c'_2) = a(\frac{n_c}{n} - \frac{m_c}{m}) + (a-1)\frac{m'_1}{m} - \frac{n'_1}{n} + (a-1)\frac{n'_2}{n} - \frac{m'_2}{m}$

de $\hat{D}(c'_1,c) \leq D(c_1,c)$ on déduit $(a-1)\frac{m'_1}{m} - \frac{n'_1}{n} \leq (a-1)\frac{m_1}{m} - \frac{n_1}{n}$

et de $\hat{D}(c,c'_2) \leq D(c,c_2)$ on déduit $(a-1)\frac{n'_2}{n} - \frac{m'_2}{m} \leq (a-1)\frac{n_2}{n} - \frac{m_2}{m}$

d'où $\hat{D}(c'_1,c'_2) \leq \hat{D}(c_1,c_2)$ et donc $(c'_1,c'_2) = (c_1,c_2)$

Cette dernière proposition permet de simplifier l'algorithme de recherche du couple $(c_1, c_2)$ optimal :

On cherche d'abord le point c tel que :

$$\sup_z |\hat{F}_1(z) - \hat{F}_2(z)| = |F_1(c) - F_2(c)|$$

Cela nous permet de déterminer les populations inférieure et supérieure.

$c_1$ est alors obtenu par maximisation de $D(z_1, c)$ et $c_2$ par maximisation de $D(c, z_2)$.

Cet algorithme nécessite 2 N investigations (N étant la taille de l'échantillon) au lieu de :

$$\frac{N^2 + N}{2}$$

si l'on avait dû examiner tous les couples $(c_1, c_2)$ avec $c_1 \leq c_2$.

Conclusion :

Le coût de l'algorithme reste linéaire par rapport à la taille de l'échantillon.

5.6. - Exemples d'applications.

Sans illustrer cette méthode et la comparer à l'algorithme de Friedman, nous présentons deux applications sur des données réelles qui correspondent à des situations limites :

Dans la première application, les deux formes à reconnaître sont bien séparées. Dans la deuxième application les deux formes à reconnaître sont assez mélangées.

APPLICATION 1

Il s'agit d'un problème de reconnaissance de spectres de machines tournantes tirées de [Cel 78].

Au vu de leurs spectres de fréquence discrétisés on doit discerner si la machine est bonne ou présente un défaut de fabrication.

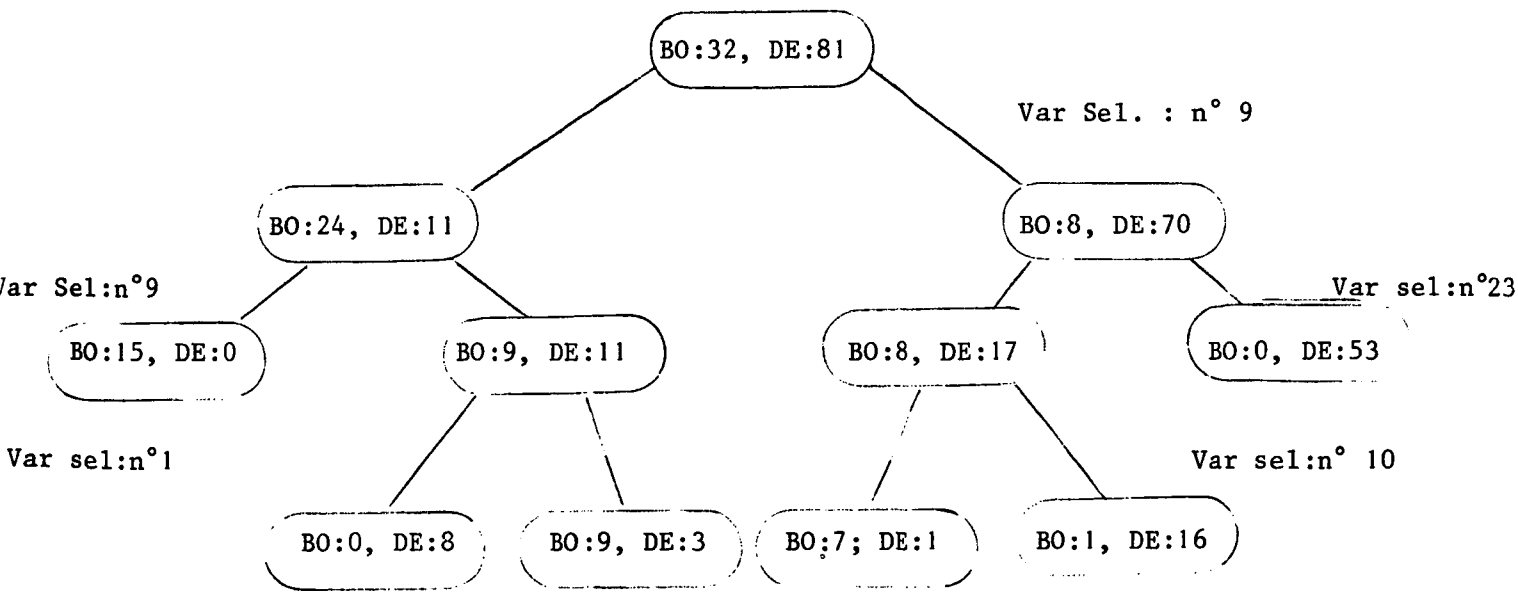
Chaque spectre est caractérisé par 24 paramètres quantitatifs exprimés en décibels.

On dispose d'un échantillon de 113 machines dont 32 sont bonnes et 86 présentent des défauts.

Pour ce problème nous avons appliqué les deux méthodes qui conduisent respectivement à un arbre de décision binaire et à un arbre de décision ternaire.

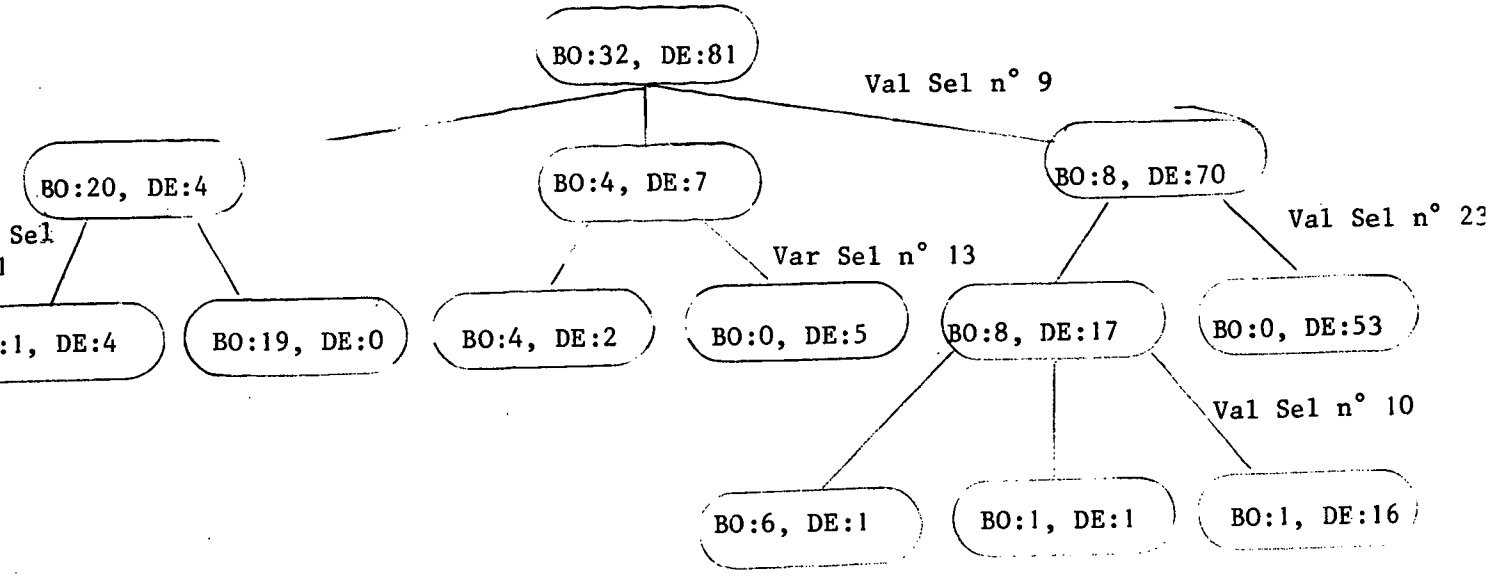
Dans la représentation des arbres, on donne, pour chaque segment, la répartition des deux formes (BO : machines bonnes, DE : machines avec défauts).

ARBRE DE DECISION BINAIRE




Il y a 4 objets mal classés : 1 du groupe BO et 3 du groupe DE.

ARBRE DE DÉCISION TERNAIRE




Il y a 5 objets mal classés : 2 du groupe BO, 3 du groupe DE et 2 objets non classés : 1 du groupe BO et 1 du groupe DE.

Cet exemple montre que dans le cas où les groupes sont bien séparés, il y a très peu de différences entre les deux méthodes ; ainsi, pour les données de Fisher les résultats sont les mêmes.

L'arbre ternaire a très peu de branches d'indécision, les variables explicatives sélectionnées sont pratiquement les mêmes et les résultats analogues.

Dans ce cas, l'arbre binaire est préférable. Il fournit d'ailleurs ici des résultats meilleurs.

APPLICATION 2.

Il s'agit d'un problème d'aide au diagnostic en médecine tiré de [Lec 77].

Nous devons discriminer les personnes bien portantes des personnes présentant une hépatite virale à l'aide de 5 paramètres quantitatifs qui sont les protéines sériques suivantes : albumine, orosmucoïde, IgG, IgN, IgA.

On dispose d'un échantillon de 549 patients : 316 sont bien portants et 233 ont une hépatite virale.

Ces deux groupes de patients sont assez mélangés. Là aussi, nous avons appliqué les deux méthodes. Nous ne reproduisons pas ici les arbres de décision qui présentent de nombreux niveaux et prennent trop de place. Les résultats finaux sont les suivants :

Pour l'arbre de décision binaire : il y a 65 mal classés sur 549 : 25 parmi les 316 bien portants, 40 parmi les 233 malades.

Pour l'arbre de décision ternaire : il y a 26 mal classés : 16 parmi les biens portants, 11 parmi les malades.

Il y a 126 individus non classés : 98 parmi les biens portants et 28 parmi les malades.

Les arbres ont des structures différentes ; de chaque noeud de l'arbre ternaire, il part systématiquement 3 branches.

L'arbre ternaire donne des pourcentages de mal classés substanciellement inférieurs en particulier pour les patients ayant une hépatite virale (11 mal classés contre 40).

Le prix payé est le suivant : 126 individus ne sont pas classés. Ils devront être affectés soit directement par le médecin, soit automatiquement à l'aide d'autres variables. Cette opération représente évidemment un coût supplémentaire.

[Pa 72] PATRICK E. "Fundamentals of pattern recognition" Prentice Hall, 1972.

[Pa Fi 70] PATRICK E., FISHER F.P. "Generalized k nearest neighbor recisear rule" Journal Information and Control, Vol. 16, N° 2, pp. 128-152, (Avril 1970).

[Sto 54] STOLLER D.S. "Univariate two-population distribution free discrimination". JASA 1954.

[Wal 44] WALD A. "On a statistical problem arising in the classification of an individual into one or two groups". Ann. Math. Statistics, Vol. 15, pp. 145-162. (1944).

[Wal 50] WALD A. "Statistical decision functions". Wiley, 1950.

[Whi 58] WHITTLE P. "On the smoothing of probability density functions" Journal of Royal Statistical. Sec. B., Vol. 20, pp. 334-343, 1958.

L'arbre de décision ternaire induit donc une procédure de reconnaissance séquentielle et un dialogue homme-machine. Par la même, elle donnera de meilleurs résultats dans les cas délicats.

## BIBLIOGRAPHIE


[And 57] ANDERSON T.W. "An introduction to multivariate statistical analysis" Wiley ; 1957.

[Cel 78] GELEUX G. Thèse de 3ème cycle ; 1978, Université Paris 6.

[Che 73] CHEN C.H. "Statistical pattern recognition". Hayden Boock Company, 1973.

[Co Ha 67] COVER T.M., HART P.E. "Nearest neighboor pattern classification" IEEE Trans. Information Theory. Vol. IT 13, n° 1, pp. 21-27 (Janvier 1967).

[Fi Ho 51] FIX E. HODGE J.L. "Discrimineary analysis ; non parametric discrimination consisteny properties". USAF School of aviation medecine project number 21.49.09. Randolp Field Texas (Février 1951).

[Fri 77] FRIEDMAN J.H. "A recursive partitionary decision rule for non parametric classification" IEEE Trans. Computer, pp. 404-408 (Avril 1977).

[Go Oh 78] GORDON L., OHLSEN R.A. "Asumptotically efficient solutions to the classification problem" Annals of Statisties. Vol. 6, n° 3, 1978.

[He Fu 69] HENRICHON E.G., FU K.S. "A non parametric partitionning procedure for pattern classification". IEEE Trans. Computer, Vol. C18, pp. 614-624. (Juin 1969).

[Ken 66] KENDALL M.G. "Multivariate analysis" ed. KRISHNAIAN, 1966.

[Le Sa 77] LECHEVALLIER Y., SANDOR G. "Découpage optimal de variables quantitatives et applications à la définition d'une grille de diagnostic tirée de l'étude des protéines résiques". 1ère journées Internationales d'Analyse des Données. INRIA 1977.

[Me Mi 73] MEISEL W.S., MICHALOPAULAS D.A. "A partitioning algorithm with application in pattern classification and optimization of decision trees" IEEE Trans. Computer, Vol. C22, pp. 93-103 (Janvier 1973).

[Par 62] PARZEN E. "On estimation of a probability density and mode" Ann. Math. Statistics, Vol. 33, n° 3, pp. 1065-1076, 1962.