2. MÉTHODES STATISTIQUES DE CLASSIFICATION

Résumé

Ce chapitre porte sur la classification supervisée. La première partie présente, d'une part, la formule de Bayes, et, d'autre part, la règle de décision de Bayes qui garantit le taux d'erreur de classification minimal.

La deuxième partie présente plusieurs méthodes de classification dont les performances approchent le taux d'erreur minimal sous certaines hypothèses. On distingue deux catégories de méthodes : les méthodes indirectes, qui utilisent la formule de Bayes, et les méthodes directes, qui évaluent les probabilités a posteriori *sans utiliser la formule de Bayes. Parmi ces dernières méthodes, les réseaux de neurones formels ont une grande qualité : ils sont capables de trouver la même solution que celle fournie par la formule de Bayes, sans condition particulière. En effet, on démontre que la sortie d'un réseau de neurones est une estimation des probabilités* a posteriori *d'appartenance aux classes. Pour chacune des méthodes présentées, nous expliquons, de manière théorique et sur un exemple, les différences de comportement.*

Dans le dernier paragraphe, nous utilisons d'une manière originale une méthode directe, telle que les réseaux de neurones, pour obtenir une estimation de la densité de probabilité d'appartenance d'un individu à une classe. Cette estimation sert ensuite au calcul indirect des probabilités d'appartenance aux classes par la formule de Bayes.

2.1 Introduction

Dans le chapitre précédent, nous avons présenté une vue d'ensemble des problèmes de classification supervisée, que nous avons séparés en deux groupes :

- Dans le premier, la règle de décision peut être expliquée de manière linguistique par le professeur : dans ce cas, la solution consiste en une suite d'opérations logiques.
- Pour les problèmes du deuxième groupe, la règle de décision ne peut être formalisée en termes linguistiques, et l'on a alors recours à une méthode statistique comportant un apprentissage supervisé à partir d'exemples.

Dans ce travail, nous nous intéressons à la résolution des problèmes de ce dernier groupe.

A partir de l'exemple consistant à classer les femmes et les hommes (voir chapitre précédent), nous introduisons le classifieur de Bayes (formule et règle de décision). La règle de décision de Bayes est incontournable en classification puisqu'elle fournit la limite théorique du taux d'erreur (ou inversement de réussite) d'un classifieur. En pratique, il convient d'approcher cette limite théorique : nous présenterons plusieurs méthodes qui y parviennent de manière plus ou moins efficace.

2.2 Présentation du problème

Comme nous l'avons déjà vu, la première étape de la conception d'un classifieur statistique consiste à faire classer, par un professeur, un échantillon d'individus définis par des descripteurs. Nous disposons alors :

- d'un échantillon composé de N individus,
- répartis dans un espace à P dimensions (les P variables descriptives),
- affectés à C classes.

Ainsi, les méthodes statistiques de résolution ne doivent s'appuyer que sur les coordonnées des individus dans l'espace de description afin de déterminer dans celui-ci plusieurs domaines qui correspondent aux classes.

Pour présenter le classifieur de Bayes, nous reprendrons l'exemple de la distinction des femmes et des hommes en fonction de la taille (l'espace de description est à une dimension : la taille).

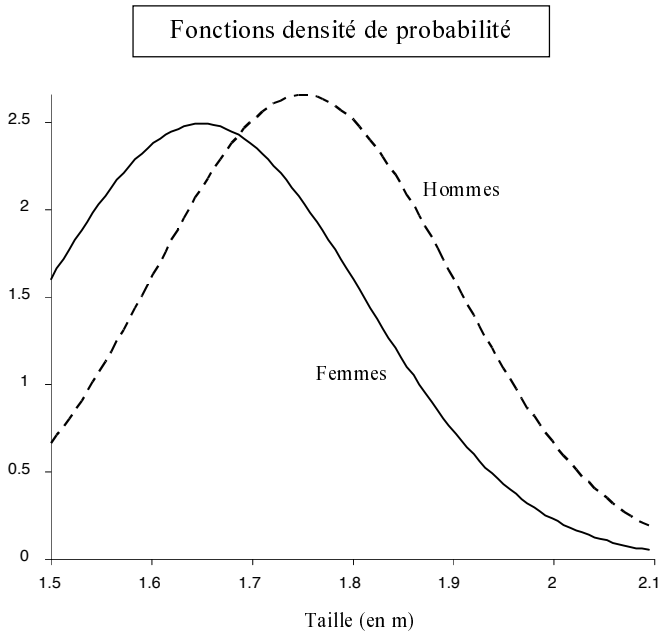


Figure 2.1 : Distribution des individus (femmes et hommes) en fonction de la taille

La figure 2.1 représente les densités de probabilité des femmes et des hommes en fonction de la taille. Nous constatons que les deux groupes d'individus[1] correspondant aux deux classes (femmes et hommes) sont décalés sans être complètement dissociés.

[1] Dans cet exemple, les classes ne sont pas disjointes. C'est le cas le plus fréquent : il est rare, dans la pratique, que l'on soit capable de trouver des variables descriptives suffisamment discriminantes pour que les classes soient complètement séparées. Dans l'exemple de classification proposé (femme/homme), choisir la taille comme unique descripteur n'est certainement pas le meilleur choix. D'autres descripteurs conduisent probablement à une représentation des formes plus adaptée.

Supposons que ces deux fonctions soient connues exactement (ce qui n'est pas le cas en général, comme nous le verrons plus loin). Alors, nous disposons d'une première grandeur caractéristique définissant les individus en fonction de leur localisation dans l'espace de description :

- $f_k(x)$: la fonction de densité de probabilité de x si la classe est k, c'est-à-dire la probabilité pour qu'un individu de la classe k soit décrit par un descripteur dont la valeur est comprise entre x et $x+\mathrm{d}x$.

  Supposons que les densités de probabilité de la taille t (en mètres) pour les classes femme et homme soient des gaussiennes :

$$f_F(t) = \frac{1}{0.16 \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\left(\frac{t-1.65}{0.16} \right)^2 \right) \text{ pour les femmes,}$$

et $$f_H(t) = \frac{1}{0.15 \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\left(\frac{t-1.75}{0.15} \right)^2 \right) \text{ pour les hommes.}$$

Quel est le sexe d'une personne mesurant 1.60 m ?

Pour répondre à cette question, nous supposons que cet individu est issu de la population entière de la France, dans laquelle nous considérons qu'il y a autant de femmes que d'hommes. A partir des densités de probabilité, la formule de Bayes donne la probabilité, dite probabilité *a posteriori*, pour qu'un individu décrit par le descripteur x appartienne à la classe k :

$$P(C_k|x) = \frac{f_k(x)}{\sum_{i=1}^{C} f_i(x)} \text{ , où } C \text{ est le nombre de classes.}$$

On a évidemment : $\sum_{k=1}^{C} P(C_k|x) = 1$

Ainsi, la formule de Bayes fournit les probabilités *a posteriori* suivantes :

$$P(F|1.60) = \frac{f_F(1.60)}{f_F(1.60) + f_H(1.60)} \approx \frac{2.38}{2.38 + 1.61} \approx 60\ \%$$

et $$P(H|1.60) = \frac{f_H(1.60)}{f_F(1.60) + f_H(1.60)} \approx \frac{1.61}{2.38 + 1.61} \approx 40\ \%$$

Avec ces valeurs, un individu provenant de la population française et mesurant 1,60 m possède 60% de chance d'être une femme (40% d'être un homme). Si on désire l'affecter à une classe, il est donc naturel de choisir celle des femmes ; ce choix (affecter l'individu à la classe pour laquelle la probabilité *a posteriori* d'appartenance est maximum), constitue la *règle de décision de Bayes*, sur laquelle nous reviendrons plus loin.

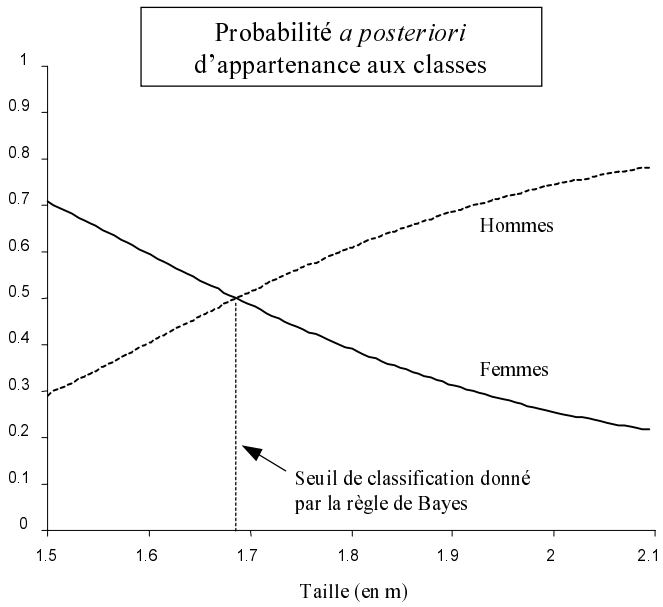


Figure 2.2 : Probabilité a posteriori *d'appartenance aux deux classes et seuil de classification (prise de décision)*

La figure 2.2 présente les probabilités *a posteriori* d'appartenance aux deux classes en fonction de la taille, calculées par la formule de Bayes. Pour un système de classification avec prise de décision binaire, l'affectation se fait en fonction du niveau de probabilité, conformément à la règle de Bayes. Pour les tailles inférieures au seuil (vers 1,68 m), le classifieur choisit la classe des femmes ; au-dessus, c'est celle des hommes. Pour une taille de 1,60 m, on retrouve bien le résultat précédent.

Compliquons un peu le problème et considérons maintenant que cet individu est un supporter d'une équipe de football.

Quel est le sexe d'un tel individu mesurant 1,60 m ?

Les distributions des individus n'ont aucune raison de changer, en revanche la proportion des femmes et des hommes dans cette population est certainement différente de celle de la population française.

Cette proportion est appelée probabilité *a priori* (elle ne dépend pas des coordonnées du point dans l'espace de description). On la note $\Pr_k$:

- $\Pr_k$: probabilité *a priori* d'appartenance à la classe k

  Pour le problème de classification femme/homme parmi les supporters, les probabilités *a priori* sont les suivantes[2] :

  $\Pr_F = 0.30$ pour les femmes,

  $\Pr_H = 0.70$ pour les hommes.

[2] Les probabilités *a priori* s'estiment d'une manière générale par un dénombrement des classes sur un échantillon de la population. Si aucune connaissance du problème n'est disponible, elles seront prises égales à $1/C$ (avec C = nombre de classes).

Pour intégrer cette nouvelle grandeur caractéristique, la formule de Bayes prend la forme plus générale suivante :

$$P(C_k|x) = \frac{\Pr_k \cdot f_k(x)}{\sum_{i=1}^{C} \Pr_i \cdot f_i(x)}$$, où C est le nombre de classes.

En tenant compte des probabilités *a priori*, on obtient les probabilités *a posteriori* :

$$P(F|1.60) \approx \frac{0.30 \cdot 2.38}{0.30 \cdot 2.38 + 0.70 \cdot 1.61} \approx 39\,\%$$

et $$P(H|1.60) \approx \frac{0.70 \cdot 1.61}{0.30 \cdot 2.38 + 0.70 \cdot 1.61} \approx 61\,\%$$

Le résultat a changé : compte tenu des probabilités *a priori* dans l'échantillon de population considéré, la probabilité que cet individu soit un homme est plus grande.

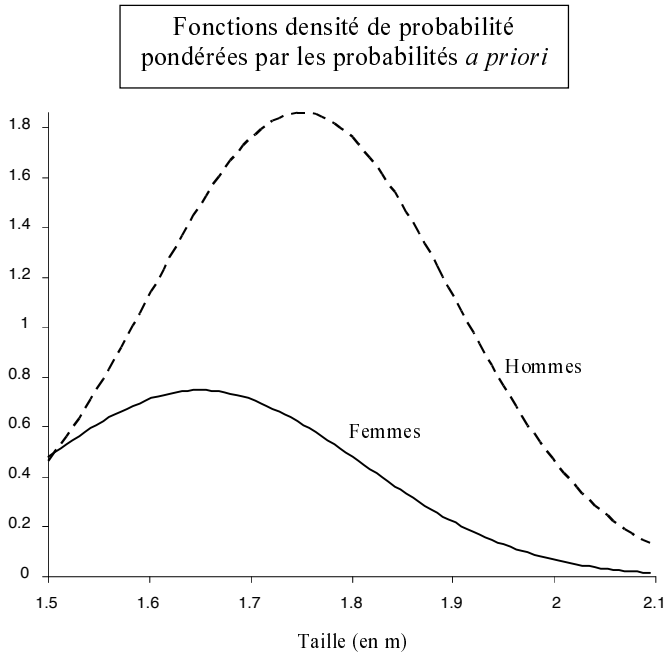


Figure 2.3 : Distribution des individus (femmes et hommes) relative aux supporters de football

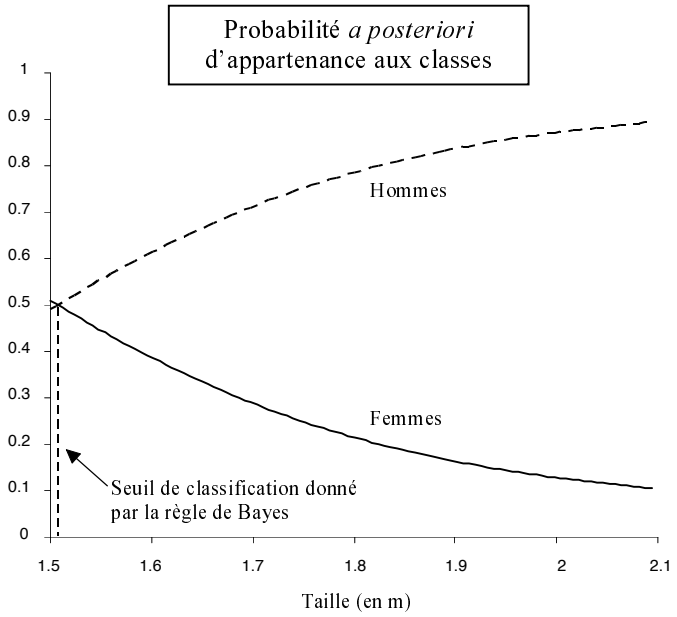


Figure 2.4 : Probabilité a posteriori *d'appartenance aux deux classes (supporters de football)*

Les figures 2.3 et 2.4 présentent les fonctions densité de probabilité et les probabilités *a posteriori* d'appartenance aux deux classes (le seuil de classification est différent de celui de la figure 2.2).

2.3 Règle de décision de Bayes

Nous rappelons l'expression mathématique de la formule de Bayes qui prend en considération la probabilité *a priori* d'apparition des individus des différentes classes et de leur distribution dans l'espace des descripteurs :

$$P(C_k|x) = \frac{\Pr_k \cdot f_k(x)}{\sum_{i=1}^{C} \Pr_i \cdot f_i(x)}$$, où C est le nombre de classes.

avec $P(C_k|\mathrm{x})$: probabilité *a posteriori* que l'individu de coordonnées x appartienne à la classe k,

$\Pr_k$: probabilité *a priori* que l'individu appartienne à la classe k,

$f_k(x)$: densité de probabilité de x si la classe est k.

Comme indiqué plus haut, la règle de décision de Bayes consiste à choisir d'affecter l'individu à la classe dont la probabilité *a posteriori* (calculée par la formule de Bayes ou par tout autre méthode) est la plus grande. On démontre [voir par exemple Duda 73] que cette décision minimise le risque d'erreur de classification.

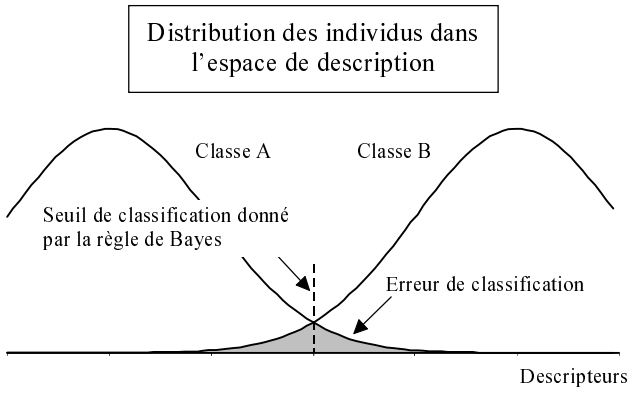
Nous pouvons également introduire le concept de coût associé à un mauvais classement. Ainsi, [Caraux 96] prend l'exemple du classement des champignons. Classer comestible un champignon toxique peut avoir des conséquences beaucoup plus dramatiques que l'inverse. Il

faut donc adopter une fonction de coût plus ou moins grande suivant le type d'erreur de classement. A la limite, si l'on désire ne prendre aucun risque, il faut classer tous les champignons comme toxiques et ne plus en consommer.

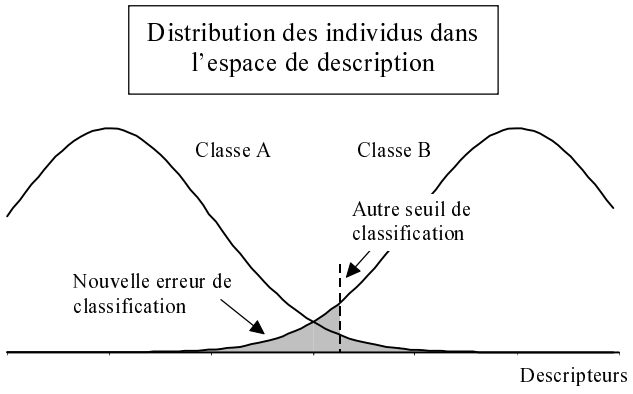
La règle de Bayes peut s'adapter à ces nouvelles conditions pour devenir une règle de décision de risque minimum (et non d'erreur de classification minimum). Nous ne présentons pas cette extension, car, dans toutes les applications de cette étude, les erreurs de classification possèdent la même importance ; nous appliquons donc la règle de décision de Bayes qui minimise l'erreur de classification.

2.4 Intérêt de la règle de décision de Bayes

Comme nous venons de l'indiquer, la règle de décision de Bayes minimise la probabilité d'erreur de classement. La figure suivante apporte une explication géométrique. [Duda 73] en donne une preuve mathématique plus rigoureuse.



a/



b/

Figure 2.5 : La règle de décision de Bayes minimise le risque d'erreur de classification

Sur la figure 2.5, les courbes représentent les fonctions densité de probabilité pondérées par les probabilités *a priori* correspondant aux deux classes A et B. De cette manière, ces

courbes sont directement reliées à la densité des individus. Sur la Figure 2.5a, le trait vertical marque le seuil de classification donné par la règle de Bayes entre les deux classes.

Quelle est l'erreur de classification commise avec un tel seuil ?

L'erreur de classification est le nombre d'exemples A classés B et inversement ; elle correspond donc à la surface grisée. Si l'on choisit un autre seuil (Figure 2.5b), nous nous apercevons que la nouvelle erreur de classification est égale à l'erreur de classification de Bayes augmentée d'une contribution positive. Elle est donc toujours supérieure à l'erreur de Bayes. Ainsi, quel que soit le seuil pris pour séparer les 2 classes, l'erreur de classification est toujours supérieure à celle trouvée avec la règle de Bayes.

En résumé, la règle de décision de Bayes constitue la limite optimale de tout système de classification. Malheureusement cette limite est théorique ; en effet, face à un problème réel, les distributions des classes (fonction de densité et probabilité *a priori*) sont inconnues, ainsi que les probabilités *a posteriori*. Les différentes méthodes mathématiques de résolution peuvent seulement en fournir des estimations.

De plus, l'erreur minimale de classification est elle-même théorique, et sa valeur est inconnue. Ce point est important, car une méthode de classification peut donner un taux de classification considéré comme insuffisant (par exemple 45% d'individus mal classés) ; dans ce cas on peut être tenté de rejeter cette méthode en la considérant comme médiocre, mais il se peut que l'erreur limite de classification de Bayes soit égale à 44% et que la méthode présente finalement d'excellentes performances. Ici, une meilleure sélection des descripteurs s'impose plus qu'une autre méthode de classification. Cette erreur théorique constitue une borne infranchissable qui représente d'une certaine manière la difficulté intrinsèque du problème.

Le paragraphe suivant propose un exemple de mise en œuvre du classifieur de Bayes. Cet exemple sera repris pour visualiser le comportement et les limites des différentes méthodes de classification présentées dans ce chapitre.

2.5 Exemple d'application du classifieur de Bayes

Un petit problème de classification à une variable descriptive est proposé pour mettre en pratique la formule et la règle de décision de Bayes. Ensuite, il sera repris pour visualiser les comportements des méthodes de classification décrites.

La figure 2.6a présente un échantillon d'individus décrits par une variable (x) à classer suivant deux classes (classe A en haut et classe B en bas).

L'échantillon comporte 1200 individus distribués de la façon suivante :

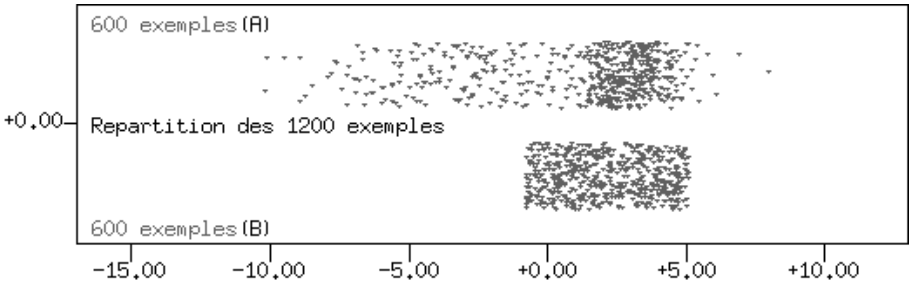
- 600 individus de classe A répartis suivant une distribution construite à partir de deux lois de Gauss.
- 600 individus de classe B répartis suivant une distribution uniforme.

La partie b/ présente les distributions des exemples suivant la variable x. Elle présente également les histogrammes de répartition des deux classes : ceci illustre la difficulté liée à

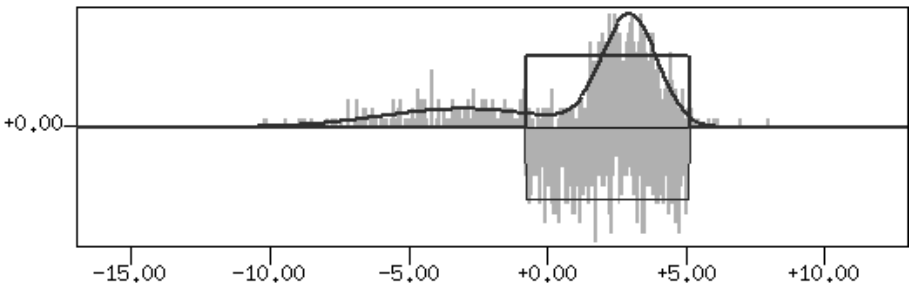
l'estimation des fonctions densité de probabilité, due au fait que cette estimation est effectuée à l'aide d'un nombre fini d'exemples.

Comme cet exemple est artificiel, nous connaissons parfaitement ses caractéristiques, c'est-à-dire les probabilités *a priori* (proportion d'individus dans chacune des classes) et les densités de probabilité (répartition des individus suivant x). On peut donc *calculer* (et non *estimer*) la probabilité *a posteriori* théorique donnée par la formule de Bayes.

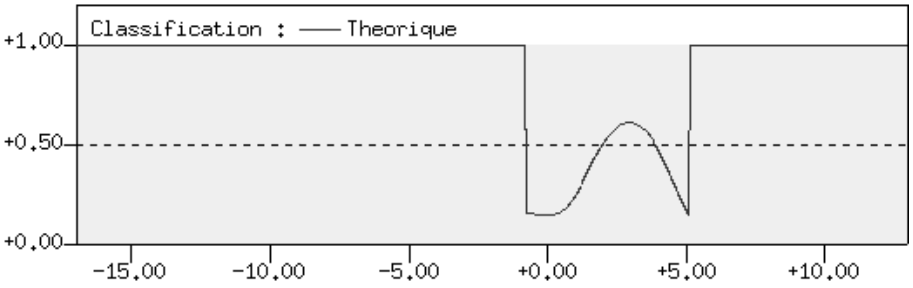
Ainsi, la partie c/ présente la probabilité *a posteriori* qu'un individu décrit par x appartienne à la classe A. Par exemple, pour $x = 0$, la fonction prend la valeur 0,15. Il y a donc 15% de chance que les individus placés autour de zéro appartiennent à la classe A ; les individus sont donc classés en B. Ainsi, tous les individus de classe A placés au voisinage de 0 sont mal classés.



a/ Visualisation des 1200 individus



b/ Fonctions et histogrammes de distribution des individus



c/ Probabilité a posteriori *d'appartenance à la classe A (TMC = 30,1%)*

Figure 2.6 : Exemple de classification à une variable descriptive

En balayant l'axe des x, nous pouvons compter le nombre d'individus mal classés, et nous obtenons ainsi un taux d'individus mal classés égal à 30,1%. Ce pourcentage est donc la

limite théorique du taux d'erreur de classification : aussi sophistiqué que soit le classifieur utilisé, il est illusoire de penser qu'il pourra réaliser un taux d'erreur de classification inférieur à 30,1%.

2.6 Méthodes indirectes de résolution

Comme nous venons de le voir, la formule de Bayes permet de déterminer les probabilités d'appartenance *a posteriori* si les densités de probabilité et les probabilités *a priori* sont connues, et la règle de décision de Bayes permet d'obtenir le taux d'erreur de classification minimum, qui est l'objectif souhaitable pour tout système de classification.

On peut donc distinguer deux groupes de méthodes de classification

- Les méthodes qui estiment les fonctions densité et les probabilités *a priori* pour ensuite calculer les probabilités *a posteriori* à l'aide de la formule de Bayes (méthodes indirectes). A l'intérieur de ce groupe, on distingue encore les méthodes paramétriques (qui font usage d'une hypothèse sur la forme analytique de la distribution) et les méthodes non paramétriques (qui ne font usage d'aucune hypothèse sur la forme de la distribution).
- Les méthodes directes, qui estiment les probabilités *a posteriori* sans faire intervenir la formule de Bayes (voir § 2.7 : Méthodes directes de résolution).

2.6.1 Estimation paramétrique des densités de probabilité

Les méthodes paramétriques consistent à faire une hypothèse concernant la forme analytique de la distribution de probabilité recherchée, et à estimer les paramètres de cette distribution à partir des données dont on dispose. En d'autres termes, à l'aide de quelques paramètres (moyenne, variance, ...) on ajuste la loi de distribution choisie par rapport aux individus à notre disposition. On obtient une estimation des paramètres, et l'on peut ensuite utiliser la forme analytique de la densité ainsi déterminée pour en déduire la densité en tout point de l'espace de représentation.

L'hypothèse la plus courante est que la répartition des individus de chacune des classes suit une loi gaussienne (loi gaussienne multidimensionnelle bien entendu). Elle conduit à la méthode appelée analyse discriminante avec une règle d'affectation probabiliste. Cette distribution "normale" des individus est la plus utilisée ; néanmoins, si notre connaissance du problème nous fait rejeter la loi de Gauss, d'autres lois peuvent la remplacer.

2.6.1.1 Analyse discriminante avec une règle d'affectation probabiliste

a/ Présentation

On rappelle l'hypothèse de distribution :

- Les individus de la classe k sont réparties suivant une loi gaussienne multidimensionnelle:

$$f_k(x) = \frac{1}{(2\pi)^{P/2}\sqrt{dét\ \Sigma_k}} \cdot \exp\left(-\tfrac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

avec Σ_k : matrice de covariance de la classe *k*,

μ_k : centre de la gaussienne de la classe *k*.

La matrice de covariance et le centre de la classe *k* sont estimés par la matrice de covariance et la moyenne des individus appartenant à la classe *k*.

Ainsi, à partir des estimations des matrices de covariance, des centres des gaussiennes (pour chacune des classes) et des probabilités *a priori*, on calcule (formule de Bayes) les probabilités *a posteriori* d'appartenance aux classes. Il ne reste plus qu'à choisir la classe qui obtient la plus grande probabilité *a posteriori*. La frontière de séparation est donc déterminée par l'ensemble des points pour lesquels les probabilités *a posteriori* sont égales.

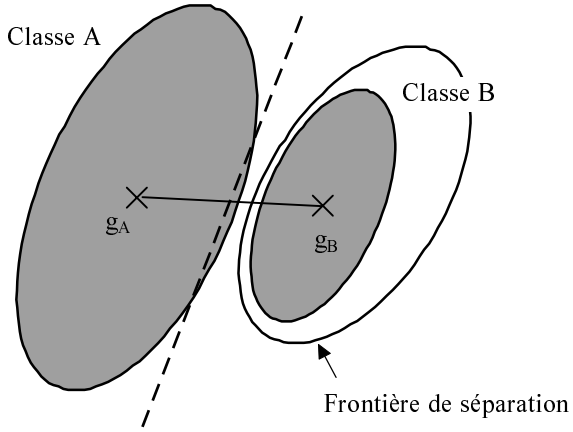


Figure 2.7 : Frontière de séparation
(analyse discriminante avec une règle d'affectation probabiliste)

Sur la figure 2.7, on remarque que la frontière de séparation prend en considération la différence de dispersion des classes. Le trait pointillé matérialise la frontière obtenue avec la règle d'affectation géométrique (voir paragraphe 2.6.1.2).

b/ Comportement

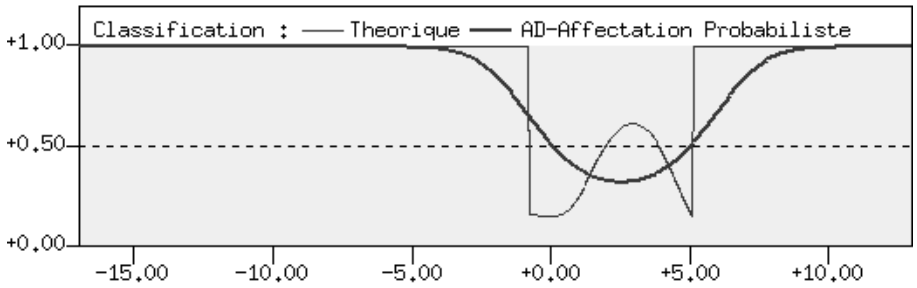


Figure 2.8 : Analyse discriminante avec une règle d'affectation probabiliste (TMC = 43,4%)

Sur la figure 2.8, nous représentons (en gras) la fonction de classification obtenue par l'analyse discriminante avec une règle d'affectation probabiliste en faisant l'hypothèse de

distribution gaussienne. Comme l'hypothèse n'est pas vérifiée, le taux d'individus mal classés (43,4%) est très supérieur à celui de Bayes (30,1%).

c/ Discussion

Cette forme de l'analyse discriminante peut sembler *a priori* très séduisante. Malheureusement même si les hypothèses sont vérifiées, les estimations des différentes matrices sont effectuées à partir des exemples disponibles, qui peuvent être peu nombreux. Elles sont donc très sensibles aux exemples marginaux.

De façon pratique, on préfère ajouter quelques hypothèses qui conduisent à l'analyse discriminante avec une règle d'affectation géométrique.

2.6.1.2 Analyse discriminante avec une règle d'affectation géométrique

a/ Présentation

C'est la forme la plus simple de l'analyse discriminante. L'hypothèse de départ est complétée par les suivantes pour garantir la convergence vers la règle de Bayes :

- Les individus de la classe k sont répartis suivant une loi gaussienne (loi de Gauss multidimensionnelle) déterminée par :

$$f_k(x) = \frac{1}{(2\pi)^{P/2}\sqrt{dét\ \Sigma_k}} \cdot \exp\left(-\tfrac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

avec Σ_k : matrice de covariance des individus de la classe *k*,

μ_k : centre de la Gaussienne de la classe *k*.

- Les différentes matrices de covariance sont identiques :

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_C = \Sigma$$

- Les probabilités *a priori* des classes sont, elles aussi, identiques :

$$\mathrm{Pr}_1 = \mathrm{Pr}_2 = \cdots = \mathrm{Pr}_C = 1/C$$

Dans ce cas, pour classer un nouvel exemple, l'analyse discriminante avec affectation géométrique calcule la distance (métrique de Mahalanobis) entre cet exemple et les différents centres de gravités des classes, et affecte à cet exemple la classe correspondant à la plus petite distance. La distance de Mahalanobis (notée Δ_Σ) est donc définie globalement dans l'espace de description par la matrice de covariance des individus :

$$\Delta_\Sigma^2(u_1, u_2) = (u_1 - u_2)^T\ \Sigma^{-1}\ (u_1 - u_2)$$

avec u_1 *et* u_2 : 2 vecteurs dans l'espace de description.

Cette méthode est dite géométrique car elle ne tient compte que de l'éloignement de l'exemple considéré aux centres de gravité : elle revient à découper l'espace par les hyperplans médiateurs des segments joignant les centres de gravité (au sens de la métrique utilisée).

Dans le cas de la classification à deux classes, on introduit *la fonction discriminante de Fisher* [Fisher 36] qui est donnée par :

$$w = e^T \Sigma^{-1} (\mu_1 - \mu_2)$$

avec w : valeur de la fonction discriminante de Fisher au point de coordonnées e,

et μ_k : centre de la gaussienne de la classe k.

Ainsi, on affectera l'observation e à la classe 1 si :

$$w = e^T \Sigma^{-1} (\mu_1 - \mu_2) > \frac{1}{2} (\mu_1 + \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$$

La figure 2.9 montre un exemple de classification à deux classes :

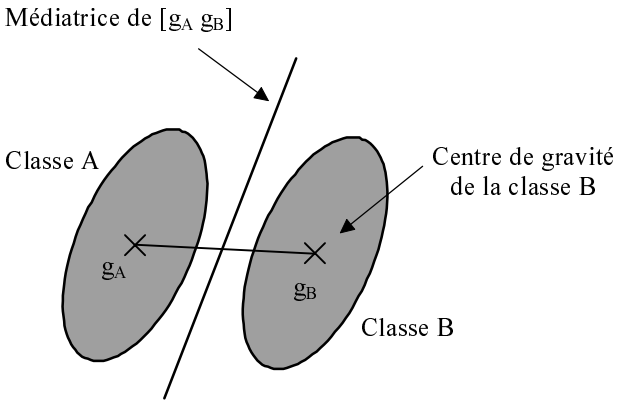


Figure 2.9 : Frontière de séparation
(analyse discriminante avec une règle d'affectation géométrique)

Au sens de la métrique de Mahalanobis, la frontière entre les deux classes (A et B) est bien la médiatrice du segment [$g_A g_B$].

Cette méthode de classification est très simple à mettre en œuvre, car elle sépare les classes suivant des hyperplans (fonctions linéaires), malheureusement le résultat obtenu est rarement (voire jamais) celui que l'on obtiendrait par le classifieur de Bayes. Ainsi, une configuration typique est celle de la figure suivante :

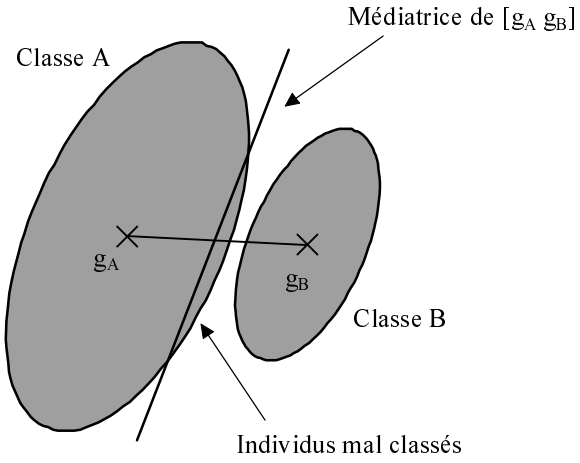


Figure 2.10 : Frontière de séparation
(analyse discriminante avec une règle d'affectation géométrique)

Ici, les individus de la classe A sont plus dispersés que ceux de la classe B. La frontière, quant à elle, n'a pas bougé par rapport à la figure 2.9 puisque les centres de gravité sont restés identiques. De nombreux individus de la classe A sont donc mal classés.

b/ Comportement

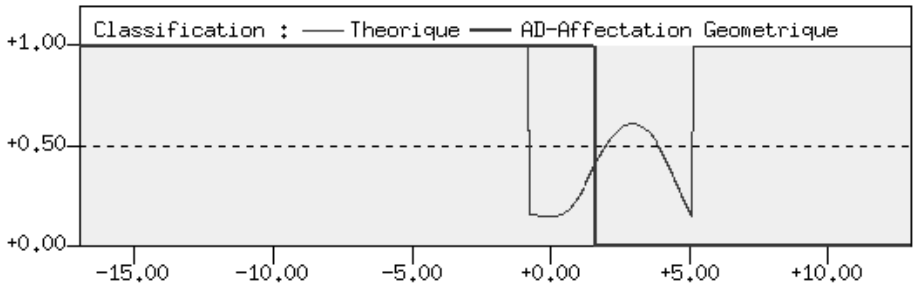


Figure 2.11 : Analyse discriminante avec une règle d'affectation géométrique (TMC = 51,7%)

La courbe en gras représente la fonction de classification obtenue par l'analyse discriminante avec une règle d'affectation géométrique. Les hypothèses ne sont pas vérifiées (la classe A est, par exemple, beaucoup plus étendue que la classe B) et le taux d'individus mal classés est de 51,7% ce qui est nettement moins bien que les 30,1% obtenus avec la règle de Bayes. Avec une telle erreur (plus de 50%), il est plus efficace de tirer au hasard la classe d'appartenance.

c/ Discussion

L'avantage de cette méthode est de faire appel à des calculs simples et de ne nécessiter que l'estimation d'une matrice de covariance à partir de tous les individus à notre disposition. De façon pratique, les conditions de convergence sont rarement vérifiées et les résultats obtenus sont généralement médiocres.

2.6.1.3 En résumé

L'analyse discriminante présente un intérêt pratique : celui d'être facile d'emploi ; les résultats qu'elle fournit sont néanmoins décevants, car les hypothèses qu'elle met en jeu ne sont généralement pas vérifiées. En effet, les distributions des individus se représentent rarement par des lois simples ; les méthodes non paramétriques peuvent apporter une solution pour tenter d'améliorer ces résultats.

2.6.2 Estimation non paramétrique des densités de probabilité

Lorsque que l'on ne peut pas faire d'hypothèse sur la distribution des individus, il faut se tourner vers des méthodes non paramétriques.

Le principe de l'estimation non paramétrique de la densité de probabilité est de délimiter une région $\boldsymbol{R}_N$ autour d'un point considéré, puis de compter le nombre d'individus dans ce volume, et enfin de déterminer la densité comme le rapport entre ce nombre (divisé par le nombre total d'individus) et le volume de la région [Parzen 62, Duda 73 et Bishop 95].

Ainsi, on obtient une estimation de la densité de probabilité avec la formule suivante :

$$\hat{p}_N(x) = \frac{k_N}{N \cdot V_N}$$

avec N : nombre d'individus de l'échantillon,

k_N : nombre d'individus dans la région $\boldsymbol{R}_N$,

V_N : volume de $\boldsymbol{R}_N$.

Trois conditions sont requises pour garantir la convergence de $\hat{p}_N(x)$ vers $p_N(x)$:

$$\lim_{N\to\infty} V_N = 0,\ \lim_{N\to\infty} k_N = \infty \text{ et } \lim_{N\to\infty} \frac{k_N}{N} = 0$$

Nous présentons deux méthodes utilisant cette propriété : les noyaux de Parzen (le volume est fixé et l'on dénombre les individus) et les k plus proches voisins (le nombre d'individus est fixé et l'on détermine le volume nécessaire pour les contenir).

2.6.2.1 Les noyaux de Parzen

a/ Présentation

Considérons un point de coordonnées x dans l'espace de description (à P dimensions) et définissons un volume (hypercube de coté h_N, avec N = nombre d'exemples) autour de ce point par :

$V_N = h_N^P$: volume de l'hypercube.

En définissant une fonction d'influence $\varphi(u)$ appelée ***noyau de Parzen*** par l'expression suivante :

$$\varphi(u)=\begin{cases}1 & \text{si } |u_j| \le 1/2 \text{ pour } j=1,...,P \\ 0 & \text{sinon}\end{cases}$$

On obtient le nombre d'exemples dans l'hypercube et l'estimation de la densité par :

$k_N(x)=\sum_{i=1}^{N}\varphi\left(\frac{x-x_i}{h_N}\right)$: nombre d'exemples k_N dans l'hypercube

L'estimateur de Parzen de la densité de probabilité est alors :

$$\hat{p}_N(x)=\frac{1}{N}\cdot\frac{1}{V_N}\cdot\sum_{i=1}^{N}\varphi\left(\frac{x-x_i}{h_N}\right)$$

L'utilisation des noyaux de Parzen requiert donc le réglage de deux "paramètres" :

- Pour assurer la convergence de l'estimateur, la dimension de la fenêtre du lissage (le volume V_N) doit répondre à deux conditions :

  $\lim_{N\to\infty} V_N = 0$ et $\lim_{N\to\infty} N\cdot V_N = \infty$

  On peut prendre par exemple : $V_N = \frac{V_1}{\sqrt{N}}$

  Le choix du volume (V_1) joue un rôle très important dans l'estimation de la densité. Si ce volume est trop grand, l'estimateur aura tendance à niveler la densité ; s'il est trop petit, l'estimateur suivra localement la présence ou non d'un exemple dans le volume.

- Le choix de la fonction noyau de Parzen est moins sensible [Caraux 96]. Toutefois, pour lisser l'estimation de la densité qui est discontinue avec la fonction $\varphi(u)$, on a souvent recours à d'autres fonctions noyau de Parzen comme par exemple le noyau gaussien, généralisé au cas multidimensionnel, dont l'expression est :

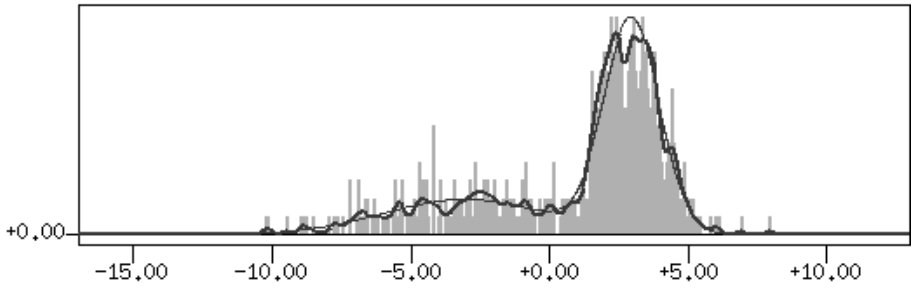
  $$K(u)=\frac{1}{(2\pi)^{P/2}\sqrt{d\acute{e}t\ \Sigma}}\cdot\exp\left(-\tfrac{1}{2}\,u^T\Sigma^{-1}u\right)$$

  où $u \in R^P$

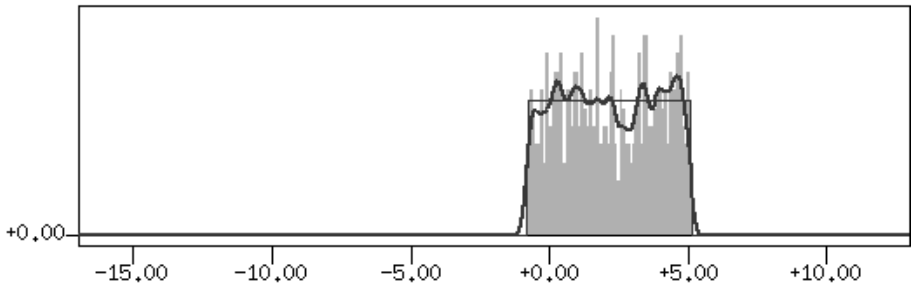
  et Σ : matrice de covariance estimée sur la description des exemples.

Dans cette étude, nous utilisons la méthode des noyaux de Parzen avec un noyau gaussien et un volume déterminé par $V_N = V_1/\sqrt{N}$.

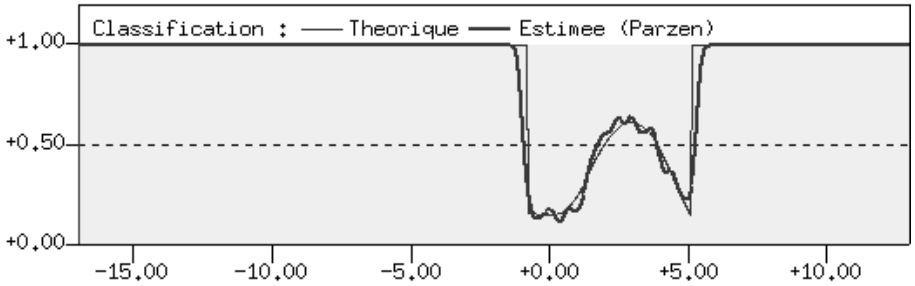
b/ Comportement



a/ Estimation de la densité de la classe A



b/ Estimation de la densité de la classe B



c/ Calcul de la probabilité a posteriori

Figure 2.12 : Noyaux de Parzen (TMC = 30,8%)

Sur la figure ci-dessus, les parties a/ et b/ présentent les estimations (en gras) des fonctions de densité de probabilité (en fin) obtenues avec les noyaux de Parzen (avec $V_1 = 4$). Sur la dernière figure, la courbe en gras représente la probabilité *a posteriori* d'appartenance à la classe A obtenue à partir des estimations des densités. On peut remarquer que les estimations des densités sont éloignées des densités théoriques, et ceci même avec beaucoup d'exemples (1200 individus) et une valeur de V_1 optimisée (elle correspond au meilleur taux de classification obtenu en utilisant cette méthode : 30,8%).

c/ Discussion

En pratique, la puissance de la méthode des noyaux de Parzen provient de sa généralité (pas d'hypothèse de distribution). Néanmoins, cette puissance se paie par un nombre d'exemples nécessaire à une bonne estimation qui croît de façon exponentielle avec la dimension.

2.6.2.2 Les k plus proches voisins

a/ Présentation

Avec les k plus proches voisins le nombre d'individus est fixé, c'est la région $\boldsymbol{R}_N$ qui grossit jusqu'à contenir les k_N individus. L'estimation de la densité de probabilité s'obtient une nouvelle fois par la relation :

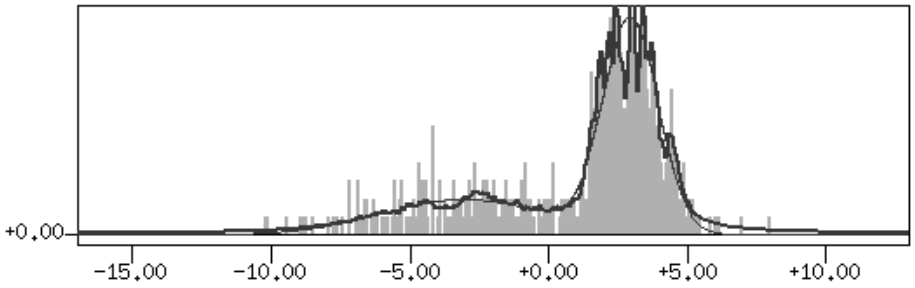
$$\hat{p}_N(x) = \frac{k_N}{N \cdot V_N}$$

L'utilisation des k plus proches voisins requiert donc le réglage du seul "paramètre" k_N. Pour assurer la convergence de l'estimateur, k_N doit répondre à deux conditions :

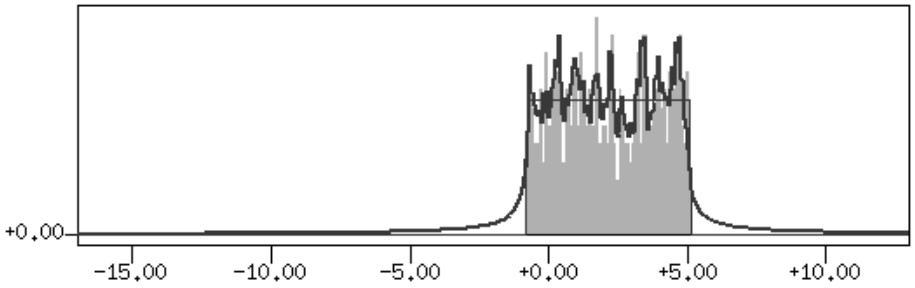
$$\lim_{N\to\infty} k_N = \infty \text{ et } \lim_{N\to\infty} \frac{k_N}{N} = 0$$

On peut prendre par exemple : $k_N = k_1 \cdot \sqrt{N}$

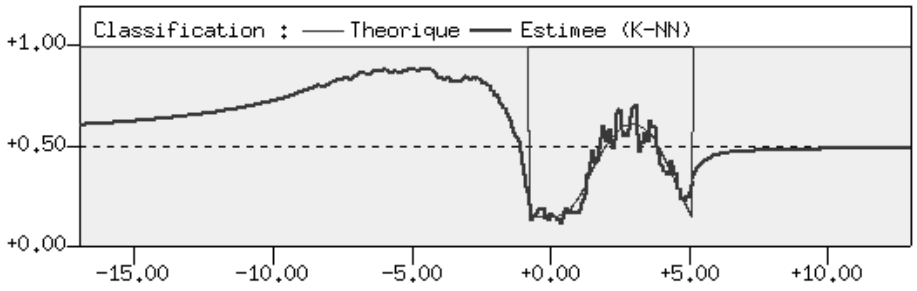
b/ Comportement



a/ Estimation de la densité de la classe A



b/ Estimation de la densité de la classe B



c/ Calcul de la probabilité a posteriori

Figure 2.13 : k plus proches voisins (k = 24 et TMC = 31,6%)

Cette figure porte sur les mêmes données que la figure 2.12, qui illustre l'estimation de densités de probabilité par la méthode des noyaux de Parzen. Ici, les estimations des densités sont obtenues avec les k plus proches voisins (avec $k_1 = 1$ et $N = 600$ et $k = k_1 \cdot \sqrt{N} = 24$). Malgré un grand nombre d'exemples et une valeur de k_1 optimisée (elle correspond au meilleur taux de classification obtenu en utilisant cette méthode : 31,6%), les estimations des densités sont assez éloignées des densités théoriques.

c/ Discussion

Là encore, la puissance de cette méthode provient de sa généralité (pas d'hypothèse de distribution) ; et là encore, cette puissance se paie par un nombre d'exemples nécessaire qui croît de façon exponentielle avec la dimension.

2.6.2.3 En résumé

Les méthodes non paramétriques d'estimation de densité de probabilité sont toutes confrontées à la "malédiction de la dimensionnalité" : le nombre d'exemples nécessaires croît de façon exponentielle avec la dimension de l'espace de description [Bellman 61]. Cet argument restreint considérablement leur champ d'application [Duda 73].

2.6.3 Classification à C classes

Dans le cas des méthodes indirectes (et contrairement aux méthodes directes, comme nous le verrons au paragraphe 2.7.2), la résolution des problèmes à C classes se fait par extension simple des méthodes de classification à deux classes. En effet, il suffit d'appliquer la formule de Bayes avec les estimations des fonctions de densité de probabilité et les estimations des probabilités *a priori*. C'est un point important, car la modification des individus d'une seule classe ne perturbe pas les estimations des caractéristiques des autres classes. Nous reviendrons sur ce point au paragraphe 2.8.

2.6.4 En résumé

D'une manière générale, il est difficile de s'attaquer à un problème sans posséder quelques connaissances *a priori* sur celui-ci. Le fait de connaître la loi de répartition des individus diminue considérablement le nombre de paramètres à ajuster et conduit à des résultats bien meilleurs. Pour cette raison, les méthodes non paramétriques d'estimation des

densités de probabilité sont très souvent moins efficaces que les méthodes paramétriques ; néanmoins, dans la pratique, il est fréquent qu'elles soient les seules utilisables.

2.7 Méthodes directes de résolution

Comme nous l'avons vu, il est possible d'estimer directement les probabilités *a posteriori* d'appartenance aux classes sans passer par l'intermédiaire des probabilités *a priori* des classes et des densités de probabilité des exemples. En effet, nous allons voir qu'il est possible de réaliser une telle estimation par minimisation d'une fonction de coût.

2.7.1 Propriété

Les méthodes directes de classification reposent sur le principe suivant :

> Soit $F(x)$ la fonction qui vaut 1 si l'individu décrit par le vecteur x appartient à la classe A, et qui vaut 0 s'il appartient à la classe B. *L'approximation au sens des moindres carrés de la fonction* $F(x)$ *constitue une estimation de* $P(\mathrm{A}|x)$ où $P(\mathrm{A}|x)$ est la probabilité *a posteriori* que l'individu de vecteur de coordonnées x appartienne à la classe A.

Cette propriété ne repose sur aucune hypothèse concernant la famille de fonctions considérée. Ainsi, tout approximateur (par exemple : droite des moindres carrés ou polynômes) permet l'estimation des probabilités *a posteriori*. Les réseaux de neurones à fonction d'activation sigmoïdale, qui présentent la propriété d'approximation universelle, sont donc de bons candidats pour réaliser cette estimation (voir Chapitre 3 : Les réseaux de neurones).

La preuve de cette propriété, ainsi que son extension à un problème de classification à C classes, sont présentées par de nombreux auteurs (voir par exemple [Bourlard 93] et [Richard 91]). Nous présentons ici, une explication plus intuitive [Rojas 96].

Plaçons-nous en un point x de l'espace de description, et délimitons autour de ce point un volume V, dans lequel nous supposons que la fonction $F(x)$ est constante. Combien d'exemples de la classe A (et aussi de B) recensons-nous à l'intérieur de V ?

Pour répondre à cette question, nous introduisons les fonctions densité de probabilité : ainsi, nous savons que la densité en un point (de coordonnées x) est proportionnelle à la fonction densité de probabilité pondérée par la probabilité *a priori*, soit :

$$d_A(x) = \mathrm{Pr}_A \cdot f_A(x) \text{ pour la classe A et}$$

$$d_B(x) = \mathrm{Pr}_B \cdot f_B(x) \text{ pour la classe B.}$$

Le nombre d'exemples des deux classes est alors donné par :

$$n_A(x) = V \cdot d_A(x) = V \cdot \mathrm{Pr}_A \cdot f_A(x) \text{ et}$$

$$n_B(x) = V \cdot d_B(x) = V \cdot \mathrm{Pr}_B \cdot f_B(x)$$

L'erreur quadratique dans le volume V est donc donnée par :

$$E_V(x) = n_A(x)\cdot\left(F(x)-1\right)^2 + n_B(x)\cdot\left(F(x)-0\right)^2$$

Cette erreur est minimale en tout point x si l'on a, pour tout x :

$\frac{\partial E_V}{\partial x} = 0$, soit en simplifiant $n_A(x)\cdot\left(F(x)-1\right) + n_B(x)\cdot F(x) = 0$

d'où $F(x) = \frac{\Pr_A \cdot f_A(x)}{\Pr_A \cdot f_A(x) + \Pr_B \cdot f_B(x)}$

On retrouve bien la probabilité *a posteriori* donnée par la règle de Bayes, donc $F(x)$ est bien une estimation de $P(\mathrm{A}|x)$.

2.7.2 Classification à C classes

Jusqu'à présent, nous n'avons considéré que des problèmes de classification à 2 classes. Le passage aux problèmes à C classes est un peu plus délicat que pour les méthodes indirectes, mais ne présente pas de difficulté majeure. Il existe plusieurs possibilités :

- La première consiste à effectuer C classifications à 2 classes en ne s'intéressant qu'à l'appartenance ou non à une classe Ainsi, pour un problème à 4 classes (A, B, C et D), on résout les 4 sous-problèmes suivants :

  - Estimation de la probabilité *a posteriori* d'appartenance à la classe A pour un individu de coordonnées x : $P(\mathrm{A}|x) = P_{\mathrm{A}}(x)$

    Pour cela, les individus de la classe A prenne la valeur désirée 1 et les autres individus prennent la valeur désirée 0[3].

  - On procède de même pour les autres sous-problèmes : $P(\mathrm{B}|x)$, $P(\mathrm{C}|x)$, et $P(\mathrm{D}|x)$.

  L'affectation est déterminée par la probabilité maximale. Notons que ces 4 probabilités ne sont pas indépendantes[4], il faut que la somme soit égale à un. On peut donc économiser un sous-problème en prenant :

  $P(\mathrm{D}|x) = 1 - P(\mathrm{A}|x) - P(\mathrm{B}|x) - P(\mathrm{C}|x)$

  Cette dernière méthode peut être numériquement très imprécise, notamment si la probabilité $P(\mathrm{D}|x)$ est petite.

- Un autre procédé consiste à décomposer le problème à C classes en C(C-1)/2 sous-problèmes à 2 classes [Knerr 92 et Price 96]. Ainsi, on résout les C(C-1)/2 sous-problèmes suivants :

  On estime la probabilité *a posteriori* que l'individu de cordonnées x appartienne à la classe A sachant qu'il appartient à la classe A ou B, notée $P_{\mathrm{AB}}(x)$. Cette

---

[3] On réalise ainsi un codage "un parmi C", appelé aussi codage "grand-mère" dans le domaine des réseaux de neurones.

[4] D'une manière pratique, lorsque l'on a obtenu une estimation de ces 4 probabilités, il est courant que leur somme soit différente de 1. Ce ne sont que des estimations de probabilités et non les probabilités *a posteriori*.

estimation est obtenue en donnant la valeur désirée 1 aux individus de la classe A et 0 aux individus de la classe B.

Naturellement, on a : $P_{BA}(x) = 1 - P_{AB}(x)$

- On procède de même pour les autres sous-problèmes : $P_{AC}(x)$, $P_{AD}(x)$, $P_{BC}(x)$, $P_{BD}(x)$ et $P_{CD}(x)$

Il ne reste plus qu'à combiner ces probabilités pour obtenir les probabilités *a posteriori* d'appartenance aux classes à l'aide la formule :

$$P(I|x) = \frac{1}{2 - C + \sum_{\substack{J=1 \\ J \neq I}}^{C} \frac{1}{P_{IJ}(x)}}$$

[Price 96] en donne une démonstration fondée sur le calcul de la probabilité de l'union d'événements non indépendants [Koroliouk 83]. Ici, notre démonstration ne fait appel qu'à la formule de Bayes :

$$P(I|x) = \frac{\Pr_I \cdot f_I(x)}{\sum_{J=1}^{C} \Pr_J \cdot f_J(x)} = \frac{Pc_I(x)}{\sum_{J=1}^{C} Pc_J(x)} = \frac{1}{\sum_{J=1}^{C} \frac{Pc_J(x)}{Pc_I(x)}} = \frac{1}{1 + \sum_{\substack{J=1 \\ J \neq I}}^{C} \frac{Pc_J(x)}{Pc_I(x)}}$$

$$= \frac{1}{1 + \sum_{\substack{J=1 \\ J \neq I}}^{C} \frac{Pc_J(x) + Pc_I(x)}{Pc_I(x)} - (C-1)} = \frac{1}{2 - C + \sum_{\substack{J=1 \\ J \neq I}}^{C} \frac{1}{P_{IJ}(x)}}$$

avec $Pc_I(x)$ = Probabilité conditionnelle (fonction de densité pondérée par la probabilité *a priori*).

Il faut encore une fois remarquer que les probabilités *a posteriori* des sous-problèmes ($P_{IJ}(x)$) ne sont pas indépendantes. D'ailleurs [Refregier 90] et [Monrocq 94] se contentent d'estimer seulement (C-1) probabilités différentes. [Price 96] apporte des éléments de comparaison entre ces méthodes : lorsque le traitement des C(C-1)/2 sous-problèmes est possible, le fait de travailler avec plus de probabilités 2 à 2 que le minimum nécessaire, rend le calcul des probabilités *a posteriori* plus fiable.

Dans cette étude, le nombre de classes est toujours très réduit (entre 3 et 5). nous utilisons donc toujours la formule précédente avec une décomposition en C(C-1)/2 sous-problèmes à 2 classes.

2.7.3 Modification des probabilités *a priori*

Avec les méthodes directes, l'estimation des probabilités *a posteriori* est évaluée à partir du seul échantillon des exemples d'apprentissage. Cependant la proportion des exemples de l'ensemble d'apprentissage peut être différente des probabilités *a priori* des classes : par

exemple, dans un problème de détection d'anomalies de fonctionnement, on peut avoir autant d'exemples d'anomalies que d'exemples de fonctionnement normal dans l'ensemble d'apprentissage, alors que, dans la réalité, les situations anormales sont beaucoup plus rares que les situations normales. Appelons $P_A(x)$ et $P_B(x)$ les probabilités *a posteriori* estimées à l'aide de la formule de Bayes à partir d'un échantillon où les probabilités *a priori* des classes sont Pr_A et Pr_B ; on sait que, en réalité, les probabilités *a priori* de ces classes sont Pr'_A et Pr'_B.

Il est possible de trouver la valeur estimée des nouvelles probabilités *a posteriori* par la relation :

$$P'_A(x) = \frac{Pr'_A}{Pr_A} P_A(x) \cdot \frac{1}{\frac{Pr'_A}{Pr_A} P_A(x) + \frac{Pr'_B}{Pr_B} P_B(x)}$$

et plus généralement :

$$P'_I(x) = \frac{Pr'_I}{Pr_I} P_I(x) \cdot \frac{1}{\sum_{J=1}^{C} \frac{Pr'_J}{Pr_J} P_J(x)}$$

Reprenons le problème de classification des femmes et des hommes : grâce à cette relation, on peut passer des probabilités *a posteriori* estimées à partir de la population française, à celle des supporters de football.

Ainsi, les outils présentés dans les deux derniers paragraphes permettent d'utiliser les méthodes directes pour résoudre n'importe quel type de problème de classification (à C classes, avec changement des probabilités *a priori*).

2.7.4 Réseaux de neurones

a/ Présentation

Comme nous le rappellerons dans le chapitre 3, les réseaux de neurones possèdent la propriété d'approcher de façon parcimonieuse n'importe quelle fonction bornée (une probabilité est bornée car comprise entre 0 et 1). Nous verrons d'autre part que les paramètres des réseaux de neurones sont estimés par minimisation d'un critère de moindres carrés[5], dont nous avons vu qu'elle permet d'estimer les probabilités *a posteriori*. L'association de ces deux propriétés fait des réseaux de neurones d'excellents candidats pour l'estimation directe des probabilités *a posteriori*.

Nous verrons en effet que ceux-ci donnent toujours de très bons résultats, car ils allient les qualités des méthodes non paramétriques (approximation universelle) à celles des méthodes paramétriques (parcimonie = peu de paramètres). [Gallinari 91] présente également les relations qui relient l'analyse discriminante aux réseaux de neurones.

b/ Comportement

---

[5] En fait, tout autre critère est aussi valable. En le minimisant, on trouve la meilleure estimation des probabilités *a posteriori* dans la famille de fonctions choisie et selon le critère choisi.

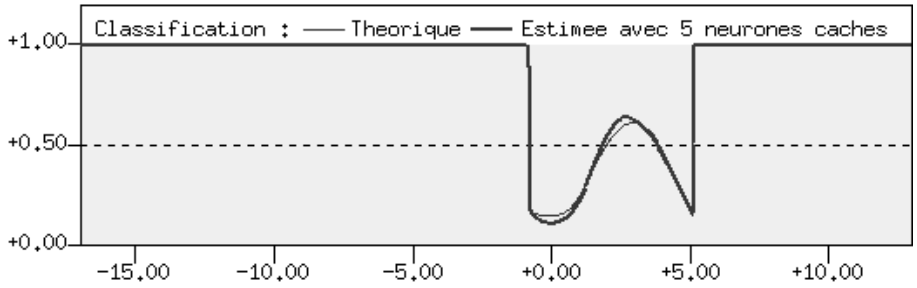


Figure 2.14 : Réseau de neurones avec 5 neurones cachés (TMC = 30,3%)

La courbe en gras représente la fonction de classification obtenue par un réseau de neurones (5 neurones cachés). Cette courbe suit de très près la courbe théorique donnée par la formule de Bayes. Le taux d'individus mal classés est de 30,3% très proche de la limite inférieure de 30,1% obtenue avec la règle de Bayes.

c/ Discussion

Le principal problème lié à l'utilisation des réseaux de neurones reste le choix de la famille de fonctions, définie par le nombre de neurones cachés d'un réseau à une couche cachée. Ce problème est d'ailleurs semblable à celui des méthodes non paramétriques : un trop grand nombre de neurones conduit à un surajustement (la fonction passe par tous les points de l'ensemble d'apprentissage, donc s'ajuste au bruit), tandis qu'un trop petit nombre de neurones ne permet pas l'ajustement. Dans le présent mémoire, nous présenterons une méthode qui permet de répondre assez rapidement et précisément à ce problème (voir chapitre 5 : La sélection de modèles). Avec de tels outils, les réseaux de neurones se révèlent être très performants.

2.7.5 En résumé

Nous avons vu que la propriété de constituer une estimation de la probabilité *a posteriori* n'appartient pas exclusivement aux réseaux de neurones. Ainsi, d'autres approximateurs usuels (tels que les polynômes) peuvent être employés ; le chapitre 3 (Les réseaux de neurones, § 3.2.2) montre que l'utilisation des réseaux de neurones est généralement plus avantageuse. Néanmoins, la méthode universelle qui donne les meilleurs résultats pour tous les problèmes n'existe pas ; pour traiter un problème particulier, la confrontation de plusieurs méthodes reste nécessaire.

Nous avons également présenté les outils nécessaires aux méthodes directes pour résoudre les problèmes de classification plus complexes, comme par exemple, ceux qui comportent C classes.

L'expérience acquise durant ce travail montre que les résultats obtenus avec un réseau de neurones bien dimensionné ne sont jamais très loin de la limite théorique de Bayes. De plus, la méthode de sélection d'architecture des réseaux de neurones (sélection des meilleures variables descriptives et des neurones cachés) permet de traiter n'importe quel problème d'une manière presque automatique.

2.8 Estimation de la densité de probabilité par une méthode originale

Dans ce paragraphe, nous présentons une méthode originale pour l'estimation d'une densité de probabilité qui utilise des réseaux de neurones. Une fois cette estimation réalisée, nous nous retrouvons dans le cadre des méthodes indirectes pour lesquelles on combine les différentes densités par la formule du classifieur de Bayes.

2.8.1 Principe

Notre problème est donc d'estimer la densité (notée $f_I(x)$) de probabilité de x si I, de densité inconnue, à partir d'un échantillon d'individus de cette classe.

L'idée est d'inverser la formule de Bayes à deux classes qui donne les probabilités *a posteriori* d'appartenance à chacune des classes si l'on connaît une bonne estimation des densités de probabilité de ces classes. En effet, nous avons vu que les réseaux de neurones permettent d'estimer directement les probabilités *a posteriori* de chaque classe. Si nous estimons directement la probabilité *a posteriori* d'une classe à l'aide d'un réseau de neurones (ou de tout autre méthode d'approximation), nous pouvons donc estimer la densité de probabilité de cette classe.

Plus précisément, le calcul de la probabilité *a posteriori* d'une classe par la formule de Bayes fait intervenir les probabilités *a priori*, et les densités de probabilité des deux classes. Créons donc une seconde classe, C, qui, contrairement à la classe I de densité inconnue $f_I(x)$, a une densité de probabilité connue $f_C(x)$ (par exemple uniforme) et une probabilité *a priori* connue ; il est facile de créer un nombre aussi grand que l'on veut d'exemples de cette classe. Disposant ainsi d'exemples des deux classes I et C, il est possible d'estimer la probabilité *a posteriori* de x si I, par exemple à l'aide d'un réseau de neurones ; on peut alors inverser la formule de Bayes pour en déduire la densité inconnue $f_I(x)$.

Soit :

N_I : nombre d'individus de la classe de densité inconnue,

$f_I(x)$: fonction de densité de probabilité **inconnue** de x si la classe est I,

N_C : nombre d'individus de la classe de densité connue,

$f_C(x)$: fonction de densité de probabilité **connue** de x si la classe est C.

Un réseau de neurones convenablement dimensionné donne une bonne estimation de la probabilité *a posteriori* de la classe inconnue. On sait d'autre part que la probabilité *a posteriori* est reliée aux densités de probabilité et aux probabilités *a priori* par la formule de Bayes :

$$P(I|x) = \frac{\frac{N_I}{N_I+N_C} \cdot f_I(x)}{\frac{N_I}{N_I+N_C} \cdot f_I(x) + \frac{N_C}{N_I+N_C} \cdot f_C(x)}$$

La densité de probabilité inconnue est donc donnée par :

$$f_I(x) = f_C(x) \cdot \frac{N_C}{N_I} \cdot \frac{P(I|x)}{1 - P(I|x)}$$

Si l'on doit résoudre un problème à deux classes inconnues, on peut ainsi estimer séparément les densités de probabilité des deux classes inconnues, puis appliquer la formule de Bayes à ces classes pour obtenir des estimations de leurs probabilités *a posteriori*.

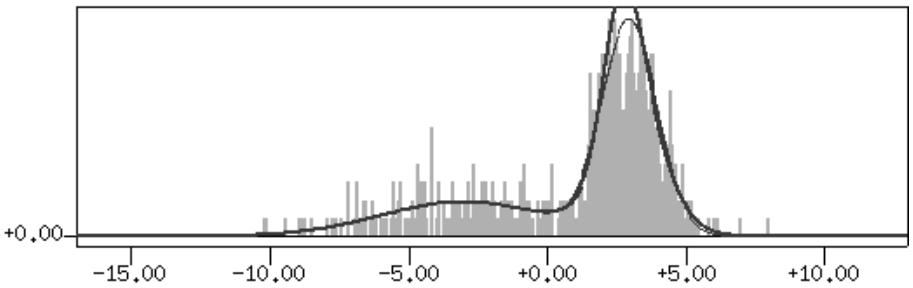
L'avantage de cette méthode est qu'elle permet de considérer les classes indépendamment les unes des autres. Ainsi, le fait de modifier l'échantillon d'une classe (par exemple en ajoutant quelques exemples apparus au dernier moment), ne demande pas un apprentissage de toutes les probabilités *a posteriori* mais seulement une nouvelle estimation de la densité de la classe. Cette remarque est particulièrement pertinente lorsque le nombre de classes est grand.

2.8.2 Utilisation de réseaux de neurones pour l'estimation de densités de probabilité

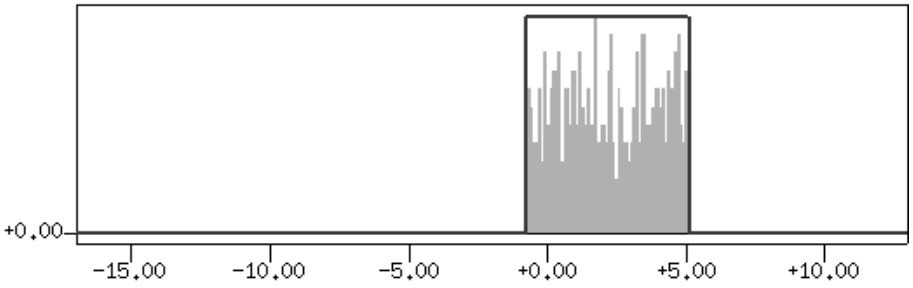
a/ Présentation

Comme nous l'avons vu, le principe de cette méthode d'estimation de la densité de probabilité est d'estimer, dans un premier temps, la probabilité *a posteriori*. Pour les mêmes raisons que celles invoquées pour les méthodes directes de classification, les réseaux de neurones sont d'excellents candidats pour l'estimation de la probabilité *a posteriori*. Tous les approximateurs (réseaux de neurones et autres) peuvent d'ailleurs être utilisés.

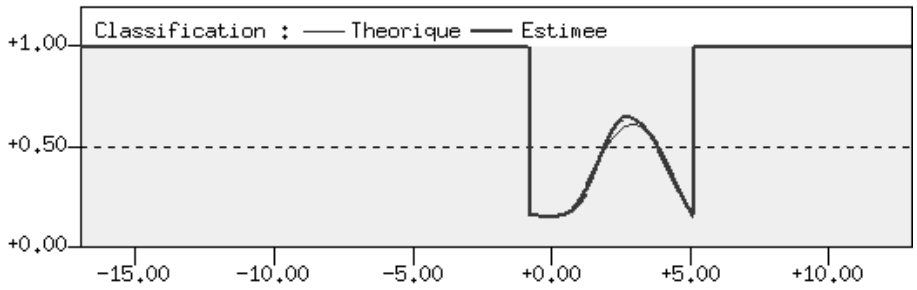
b/ Comportement



a/ Estimation de la densité de la classe A (3 neurones cachés)



b/ Estimation de la densité de la classe B (2 neurones cachés)



c/ Calcul de la probabilité a posteriori

Figure 2.15 : Estimation de la densité de probabilité par réseau de neurones (TMC = 30,2%)

Cette figure porte sur les mêmes données que la figure 2.12, qui illustre l'estimation de densités de probabilité par la méthode des noyaux de Parzen. Ici, les estimations des densités se font en utilisant les réseaux de neurones dimensionnés automatiquement avec la méthode présentée plus loin (voir chapitre 5 : La sélection de modèles). L'estimation de la probabilité *a posteriori* est très proche de la probabilité théorique et conduit à un taux d'erreur de classification de 30,2% quasiment identique à la limite de Bayes (30,1%).

c/ Discussion

Cette méthode se heurte évidemment aux difficultés habituelles d'utilisation des réseaux de neurones (choix des meilleures variables descriptives et du nombre de neurones cachés). Avec les outils de définition automatique de l'architecture d'un réseau (voir chapitre 5 : La sélection de modèles), les réseaux de neurones se révèlent très performants.

2.8.3 En résumé

Cette méthode originale de classification paraît potentiellement très puissante. En effet, elle possède les avantages des méthodes indirectes tout en conservant celles des réseaux de neurones :

- Les classes sont traitées séparément. Ainsi, une mauvaise répartition de l'échantillon d'une classe n'influence pas les estimations des probabilités d'une autre classe. De plus, si une classe est modifiée après l'apprentissage, il suffit de réajuster l'estimation de sa densité, et seulement celle-ci.
- Avec la propriété fondamentale des réseaux de neurones (approximation universelle et parcimonieuse), l'estimation de la densité nécessitera moins d'exemples que les méthodes non paramétriques (noyaux de Parzen par exemple).

[Bishop 95] présente une autre méthode d'estimation de densités de probabilité appelée "*mixture-of-experts model*". Cette méthode est fondée sur une estimation paramétrique de la densité (comme par exemple une combinaison linéaire de noyaux gaussiens). L'originalité de la méthode provient de l'estimation des paramètres de la combinaison (centres et variances des gaussiennes et coefficients de pondération), au point considéré, par un réseau de neurones. Malheureusement, on retrouve avec cette méthode, les principaux défauts des techniques mises en jeu. Dans un premier temps, il faut déterminer le nombre de noyaux nécessaires au "mixture model" ; ensuite, il faut choisir la bonne architecture du réseau de neurones.

2.9 Conclusion

Pour résoudre un problème de classification, nous disposons donc de toute une panoplie de méthodes qu'il faut utiliser en connaissant bien leur capacité et surtout leurs limitations. Dans toutes les expériences (théoriques et pratiques) que nous avons effectuées au cours de ce travail, les réseaux de neurones conduisent à de bons résultats. Bien dimensionnés, ils obtiennent toujours les meilleurs taux d'erreur de classification.

La méthode originale d'estimation des fonctions de densité par réseau de neurones semble prometteuse, car elle allie la souplesse d'utilisation des méthodes indirectes à la parcimonie des réseaux de neurones. Une comparaison plus systématique de cette méthode par rapport aux autres reste nécessaire pour la valider.