Reconnaissance des Formes et décision en Signal

Régis Lengellé
coordonnateur

version v1.0e, 1er avril 1999

Table des matières

Chapitre 1

Décision à structure imposée, contrôle de la complexité et sélection de modèle

Régis Lengellé et Cédric Richard[1]

Abstract: bla bla bla en anglais

Mots-clés : Décision statistique, sélection de modèle, contrôle de la complexité, dimension de Vapnik, compromis biais-variance.

1.1. Introduction et définitions

Dans les chapitres précédents, les connaissances nécessaires à l'élaboration d'une règle de décision étaient les lois de vraisemblance des observations conditionnellement aux M hypothèses en concurrence, dont les paramètres peuvent être connus (cas des hypothèses simples) ou inconnus (cas des hypothèses composées). Pour une classe assez large de problèmes, l'élaboration d'un modèle probabiliste adéquat n'est pas toujours chose facile. En revanche, lorsqu'une expertise des phénomènes observés est disponible, il peut être assez simple de recueillir un ensemble de données pour lesquelles un ou plusieurs experts peuvent fournir un étiquetage, c'est à dire la décision idéale pour chacune des observations. Cet ensemble de données étiquetées constitue un ensemble d'apprentissage A_N dont on supposera qu'il est exhaustif, c'est à dire qu'il reflète l'ensemble des hypothèses en concurrence.

$$A_N = \{\mathbf{X}_k, Y_k\}_{k=1,\dots,N} \tag{1.1}$$

[1] UTT, LM2S

$\mathbf{X}_k \in E$ représente l'observation k, de dimension d et $Y_k, k \in \{0, \ldots, M-1\}$ la décision associée fournie par un expert et supposée connue avec certitude. N est la taille de cet ensemble d'apprentissage. Le couple $(\mathbf{X}, Y)$ est issu de la loi conjointe : $p(\mathbf{x}, y)$ définie sur $\mathbb{R}^d \times \{0, 1\}$, mais supposée inconnue.

A chaque hypothèse est associée une classe (notée ω). Si l'hypothèse H_i est vérifiée pour l'observation $\mathbf{X}$ $(Y = i)$, on dira que $\mathbf{X} \in \omega_i$.

1.1.1. *Théorie de la décision et Reconnaissance des Formes*

En théorie statistique de la décision, l'élaboration d'une règle de décision repose sur :

- la connaissance d'un modèle probabiliste des observations
- un critère de performance

La règle de décision obtenue, appliquée à une observation $\mathbf{x}$ fournira en sortie une décision d_i (décision en faveur de l'hypothèse H_i). En fait, à l'élaboration d'une règle de décision correspond la recherche de la partition $E_1, ..., E_M$ de E qui optimise le critère de performance choisi. On prendra alors la décision d_i pour toutes les observations $\mathbf{x} \in E_i$.

Dans le cas de l'approche Reconnaissance des Formes (RdF), la connaissance d'un modèle probabiliste est remplacée par celle d'un ensemble d'apprentissage A_N. Là encore, l'élaboration d'une règle de décision consiste à proposer une partition de l'espace des observations E qui optimise le critère de performance choisi.

On distingue alors deux approches possibles :

- on souhaite fournir une décision pour toute observation à l'aide d'une fonction de l'observation, (cette fonction étant paramétrée, le processus d'optimisation portant sur ces paramètres) ; on parle alors de décision à structure imposée
- on utilise directement l'ensemble d'apprentissage pour la prise de décision et l'on parle alors de décision non paramétrique.

Ce chapitre est dédié à la décision à structure imposée.

Il faut alors garder à l'esprit que toute solution obtenue à l'aide d'un ensemble d'apprentissage sera une solution approchée. On appliquera un principe qui permet de trouver une solution approchée à l'aide d'un l'ensemble d'apprentissage (principe "d'induction"). Afin d'illustrer ce point, la solution la plus naturelle peut consister à optimiser une estimation, faite sur l'ensemble d'apprentissage, du critère de performance que l'on a choisi. Supposons que

l'on essaye d'approcher y avec une fonction $g(\mathbf{x}, \theta)$ (paramétrée par θ) en minimisant :

$$C(\theta) = \int Q(g(\mathbf{x}, \theta), y) p(\mathbf{x}, y) d\mathbf{x} dy \tag{1.2}$$

où Q représente le coût associé à un couple $(\mathbf{x}, y)$ particulier.

Dans le cas de la minimisation de l'erreur quadratique, nous obtiendrons :

$$C(\theta) = \int (y - g(\mathbf{x}, \theta))^2 p(\mathbf{x}, y) d\mathbf{x} dy \tag{1.3}$$

Dans le cas de la minimisation de la probabilité d'erreur, nous aurons :

$$C(\theta) = \int I_{g(\mathbf{x}, \theta) \neq y} p(\mathbf{x}, y) d\mathbf{x} dy \tag{1.4}$$

Le principe d'induction précédent conduira à l'optimisation de :

$$C_e(\theta) = 1/N \sum_{i=1}^{N} (y_i - g(\mathbf{x}_i, \theta))^2 \tag{1.5}$$

ou de :

$$C_e(\theta) = 1/N \sum_{i=1}^{N} I_{g(\mathbf{x}_i, \theta) \neq y_i} \tag{1.6}$$

respectivement. La loi des grands nombre garantit alors la convergence du coût empirique vers le coût théorique, lorsque la taille de la base d'apprentissage tend vers l'infini.

Maintenant que les bases sont posées, il est entendu que, dans le cas de l'approche classique de la décision ou de l'approche RdF, on ne cherchera pas nécessairement à exprimer la partition obtenue après optimisation du critère de performance, mais simplement à fournir la décision pour toute observation future $\mathbf{x}$, ce qui est équivalent $(\mathbf{x} \in E_i \Leftrightarrow d(\mathbf{x}) = d_i)$. En RdF la décision $d(\mathbf{x}) = d_i$ est notée usuellement sous la forme $\mathbf{x} \in \omega_i$.

La constitution de l'ensemble d'apprentissage nécessite de prendre quelques précautions :

– il doit être exhaustif

- il doit (si possible) refléter les probabilités a priori des classes
- il doit refléter les distributions de probabilité conditionnelle

Le premier point n'est pas toujours aisé à vérifier. En effet, si l'on se place, par exemple, dans le cadre de la décision relative à la surveillance d'un système critique au niveau de la sécurité, nous ne disposerons pas, en général, de données correspondant à ces états critiques. Il est donc important de prévoir une ou plusieurs classes supplémentaires auxquelles on affectera les observations dont on estime qu'elles ne peuvent appartenir à une et une seule des classes de l'ensemble d'apprentissage. On parlera de rejet de distance lorsque l'observation n'appartient raisonnablement à aucune classe de A_N et de rejet d'ambiguïté lorsque l'on estime que l'observation peut appartenir à plusieurs classes. Quant aux probabilités a priori, leur connaissance n'est pas critique, en particulier si l'on introduit des coûts dans le critère de performance. On peut alors, par exemple, compenser une mauvaise représentativité des classes par un choix judicieux des coûts. Rappelons toutefois que l'information statistique relative à une classe sera d'autant plus grande que l'effectif correspondant sera élevé.

Le choix de l'espace de représentation E est largement dépendant de l'application considérée et fait appel, en général, au recueil de l'expertise. Les composantes de $\mathbf{X}$ sont appelées "variables" ou "descripteurs" et sont définies en essayant, par exemple, de prendre en compte l'invariance souhaitée vis à vis de transformations attendues, ou de maximiser l'information discriminante apportée par les composantes retenues. Cette étape d'extraction d'information conditionne largement les performances du système de décision et, lorsqu'elle est bien réalisée, garantit la robustesse de l'approche Reconnaissance des Formes. Hélas, il n'existe pas de méthode garantissant l'optimalité du choix de l'espace de représentation, le meilleur étant évidemment défini par une statistique suffisante de dimension minimale, dont la détermination nécessite la connaissance des lois de probabilité conditionnelles, inconnues ici. Il sera simplement possible de comparer, a posteriori, plusieurs espaces de représentation au vu des performances obtenues, de combiner les composantes de $\mathbf{X}$ [FUK 72] ou d'en sélectionner le meilleur sous-ensemble de dimension d' avec $d' < d$ en évitant l'explosion combinatoire [NAR 77].

Dans le cas des signaux échantillonnés, les échantillons temporels sont supposés constituer une statistique suffisante (de dimension maximale) pour la prise de décision. On peut alors envisager de choisir $\mathbf{X} = (x(1), \ldots, x(d))^t$, où d représente le nombre d'échantillons de chaque observation. Ce choix peut conduire à la confrontation au fléau de la dimensionnalité [BEL 61]. En effet, la densité moyenne des observations dans E est proportionnelle à $N^{\frac{1}{d}}$. Cela exprime le fait que la taille de l'ensemble d'apprentissage doit croître exponentiellement avec la dimension de l'espace de représentation. En effet, si l'on estime, pour un problème donné dans $\mathbb{R}$, qu'un ensemble d'apprentissage de

taille 10 est suffisant, le même problème transposé dans $\mathbb{R}^q$ nécessite environ 10^q observations pour arriver à une densité comparable, ce qui devient rapidement irréaliste, même pour des valeurs faibles de q. Nous reviendrons plus loin sur ce point.

1.1.2. *Consistance de la règle de décision*

L'approche considérée ici consiste à rechercher une fonction $d(\mathbf{X})$:

$$d(\mathbf{X}) : \mathbb{R}^d \rightarrow \{0, \ldots, M-1\} \tag{1.7}$$

qui permette de fournir une décision $Y = d(\mathbf{X})$ pour toute observation future. En général, Y n'est pas une fonction déterministe de $\mathbf{X}$. Le modèle communément admis pour Y est :

$$Y = g(\mathbf{X}) + \varepsilon \tag{1.8}$$

où $g(.)$ est une fonction déterministe et ε est une variable aléatoire. La non unicité de ce modèle (on peut toujours ajouter une constante à $g(.)$ et la soustraire à ε) est levée en considérant que $E(\varepsilon) = 0$.

Afin de simplifier l'exposé, nous allons considérer dans ce chapitre le cas de la décision à hypothèses binaires (détection) qui correspond au cas $M = 1$.

Le couple $(\mathbf{X}, Y)$ est issu de la loi conjointe : $p(\mathbf{x}, y)$ définie sur $\mathbb{R}^d \times \{0, 1\}$, mais inconnue.

La probabilité d'erreur associée à d est :

$$P_e = p(d(\mathbf{X}) \neq Y) \tag{1.9}$$

La détermination de la meilleure fonction d nécessite la définition d'un critère de performance à optimiser. Si celui-ci est la probabilité d'erreur P_e, la solution optimale d^* est définie par :

$$d^* = \arg\min_d p(d(\mathbf{X}) \neq Y) \tag{1.10}$$

qui est le détecteur de Bayes avec des coûts (0,1). Celui-ci conduit à $P_e^* = p(d^*(\mathbf{X}) \neq Y)$. Dans le cas d'hypothèses binaires (2 classes ω_0 et ω_1), la règle de décision correspondante s'écrit :

$$d^*(\mathbf{x}) = \begin{cases} 1 & \text{si} \quad p(\omega_1/\mathbf{x}) \geq p(\omega_0/\mathbf{x}) \\ 0 & \text{si} \quad p(\omega_1/\mathbf{x}) < p(\omega_0/\mathbf{x}) \end{cases} \tag{1.11}$$

ou, de manière identique :

$$d^*(\mathbf{x}) = \begin{cases} 1 & \text{si} \quad p(\omega_1/\mathbf{x}) \geq 1/2 \\ 0 & \text{si} \quad p(\omega_1/\mathbf{x}) < 1/2 \end{cases} \tag{1.12}$$

Notons que l'évaluation de P_e nécessite la connaissance de $p(\mathbf{x}, y)$, inconnue ici. En fait, la seule information disponible étant A_N, le critère de performance dépendra de A_N et il en sera de même de toute fonction d obtenue. Nous la noterons $d_N(\mathbf{X}; A_N)$.

Pour un ensemble d'apprentissage A_N *donné*, cette règle va conduire à la probabilité d'erreur $P_e(d_N; A_N)$ définie par :

$$P_e(d_N; A_N) = p(d_N(\mathbf{X}; A_N) \neq Y/A_N) \tag{1.13}$$

qui est une variable aléatoire (fonction de A_N). La probabilité d'erreur associée à $d_N(\mathbf{X}; A_N)$ sera $P_e(d_N) = E(P_e(d_N; A_N))$.

Considérons maintenant le cas où la taille N de l'ensemble d'apprentissage A_N augmente. Nous pouvons espérer quelques propriétés asymptotiques de notre détecteur.

Définition 1.1.1 (Consistance) *Un détecteur est dit consistant pour une distribution $p(\mathbf{x}, y)$ donnée si*

$$\lim_{N \to \infty} E(P_e(d_N; A_N))) = P_e^*. \tag{1.14}$$

On dira qu'il est fortement consistant si

$$\lim_{N \to \infty} P_e(d_N; A_N)) = P_e^* \text{ presque sûrement} \tag{1.15}$$

Enfin, on dira qu'il est universellement (fortement) consistant s'il est (fortement) consistant pour toute distribution $p(\mathbf{x}, y)$.

Cette dernière propriété, très forte, a pu être démontrée pour la première fois en 1977 par Stone [STO 77] pour la méthode des k plus proches voisins lorsque $k(N) \to \infty$ et $k/N \to 0$.

La section suivante fait le lien entre décision à partir d'une base de données étiquetées et régression. En particulier, nous verrons que la décision est un problème plus simple à traiter que la régression.

1.2. Régression et détection

Les équations [1.10] et [1.11] illustrent le rôle central que jouent les probabilités a posteriori $p(\omega_i/\mathbf{x}), i = 1, 2$. On peut alors imaginer élaborer une règle de décision qui repose sur l'estimation $\hat{p}_N(\omega_i/\mathbf{x}) \equiv \hat{p}(\omega_i/\mathbf{x}; A_N)$ de $p(\omega_i/\mathbf{x})$. C'est l'application d'un principe d'induction qui conduit naturellement à l'utilisation de l'ensemble d'apprentissage pour estimer $p(\omega_i/\mathbf{x})$. Dans ce cas, la règle de décision $d_N(\mathbf{x}; A_N)$ s'écrit :

$$d_N(\mathbf{x}) \equiv d_N(\mathbf{x}; A_N) = \begin{cases} 1 & \text{si} \quad \hat{p}_N(\omega_1/\mathbf{x}) \geq 1/2 \\ 0 & \text{si} \quad \hat{p}_N(\omega_1/\mathbf{x}) < 1/2 \end{cases} \tag{1.16}$$

On définira $\hat{p}_N(\omega_0/\mathbf{x})$ par :

$$\hat{p}_N(\omega_0/\mathbf{x}) = 1 - \hat{p}_N(\omega_1/\mathbf{x}) \tag{1.17}$$

Avec une telle approche, pour tout estimé $\hat{p}_N(\omega_1/\mathbf{x})$ non négatif, nous avons :

Théorème 1.2.1 $P_e(d_N; A_N) - P_e^* \leq 2 \int_{\mathbb{R}^d} |p(\omega_1/\mathbf{x}) - \hat{p}_N(\omega_1/\mathbf{x})| p(\mathbf{x}) d\mathbf{x}$

Preuve :

Si, pour une valeur $\mathbf{x}$ de $\mathbb{R}^d$ nous avons

$$d_N(\mathbf{x}) = d^*(\mathbf{x}),$$

alors

$$p(d_N(\mathbf{X}) \neq Y/\mathbf{X} = \mathbf{x}) - p(d^*(\mathbf{X}) \neq Y/\mathbf{X} = \mathbf{x}) = 0$$

sinon :

$$\begin{aligned} p(d_N(\mathbf{X}) \neq Y/\mathbf{X} = \mathbf{x}) &= 1 - p(d_N(\mathbf{X};) = Y/\mathbf{X} = \mathbf{x}) \\ &= 1 - \{p(Y = 1, d_N(\mathbf{X}) = 1/\mathbf{X} = \mathbf{x}) + p(Y = 0, d_N(\mathbf{X}) = 0/\mathbf{X} = \mathbf{x})\} \\ &= 1 - \left\{\mathbf{I}_{d_N(\mathbf{x})=1} p(Y = 1/\mathbf{X} = \mathbf{x}) + \mathbf{I}_{d_N(\mathbf{x})=0} p(Y = 0/\mathbf{X} = \mathbf{x})\right\} \\ &= 1 - \left\{\mathbf{I}_{d_N(\mathbf{x})=1} p(\omega_1/\mathbf{x}) + \mathbf{I}_{d_N(\mathbf{x})=0} (1 - p(\omega_1/\mathbf{x}))\right\} \end{aligned}$$

d'où, pour tout $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned} p(d_N(\mathbf{X}) \neq Y/\mathbf{X} = \mathbf{x}) - p(d^*(\mathbf{X}) \neq Y/\mathbf{X} = \mathbf{x}) \\ = p(\omega_1/\mathbf{x})(\mathbf{I}_{d^*(\mathbf{x})=1} - \mathbf{I}_{d_N(\mathbf{x})=1}) + (1 - p(\omega_1/\mathbf{x}))(\mathbf{I}_{d^*(\mathbf{x})=0} - \mathbf{I}_{d_N(\mathbf{x})=0}) \\ = (2p(\omega_1/\mathbf{x}) - 1)(\mathbf{I}_{d^*(\mathbf{x})=1} - \mathbf{I}_{d_N(\mathbf{x})=1}) \\ = |2p(\omega_1/\mathbf{x}) - 1| \mathbf{I}_{d_N(\mathbf{x}) \neq d^*(\mathbf{x})} \end{aligned}$$

donc : $P_e(d_N; A_N) - P_e^* = \int_{\mathbb{R}^d} 2|p(\omega_1/\mathbf{x}) - 1/2|\mathbf{I}_{d_N(\mathbf{x}) \neq d^*(\mathbf{x})} p(\mathbf{x})d\mathbf{x}$

soit encore :

$P_e(d_N; A_N) - P_e^* \leq \int_{\mathbb{R}^d} 2|p(\omega_1/\mathbf{x}) - \hat{p}_N(\omega_1/\mathbf{x})|p(\mathbf{x})d\mathbf{x}$

car

$d_N(\mathbf{x}) \neq d^*(\mathbf{x}) \Rightarrow |p(\omega_1/\mathbf{x}) - \hat{p}_N(\omega_1/\mathbf{x})| \geq |p(\omega_1/\mathbf{x}) - 1/2|$

CQFD.

En remarquant que :

$\int_{\mathbb{R}^d} |p(\omega_1/\mathbf{x}) - \hat{p}_N(\omega_1/\mathbf{x})|p(\mathbf{x})d\mathbf{x} = E(|p(\omega_1/\mathbf{x}) - \hat{p}_N(\omega_1/\mathbf{x})|/A_N)$,

on peut déduire directement du théorème [1.2.1] que :

$P_e(d_N; A_N) - P_e^* \leq 2\sqrt{\int_{\mathbb{R}^d} |p(\omega_1/\mathbf{x}) - \hat{p}_N(\omega_1/\mathbf{x})|^2 p(\mathbf{x})d\mathbf{x}}$.

En conséquence, si l'on peut proposer une règle de décision $d_N(\mathbf{x})$ qui repose sur un estimateur $\hat{p}_N(\omega_1/\mathbf{x})$ tel que :

$\int_{\mathbb{R}^d} |\hat{p}_N(\omega_1/\mathbf{x}) - p(\omega_1/\mathbf{x})|^2 p(\mathbf{x})d\mathbf{x}$

$= E(|\hat{p}_N(\omega_1/\mathbf{x}) - p(\omega_1/\mathbf{x})|^2) \to 0$ lorsque $N \to \infty$

en probabilité (presque sûrement), alors cette règle de décision est consistante (fortement consistante). La consistance de la règle de décision repose donc sur celle de l'estimateur de la probabilité a posteriori $p(\omega_1/\mathbf{x})$.

Afin de faire le lien entre décision et régression, il convient de remarquer que

$E(Y/\mathbf{X} = \mathbf{x}) = 1p(Y = 1/\mathbf{x}) + 0p(Y = 0/\mathbf{x}) = p(\omega_1/\mathbf{x})$. La probabilité a posteriori $p(\omega_1/\mathbf{x})$ n'est autre que la régression de la variable Y sur les observations $\mathbf{X}$, d'où les liens très forts entre décision à partir d'une base de données étiquetées et régression, car la fonction $g(.)$ qui minimise $E(Y - g(\mathbf{X}))^2$ est $g(\mathbf{X}) = E(Y/\mathbf{X})$. En effet :

$E((Y - p(\omega_1/\mathbf{X})^2) \leq E((Y - g(\mathbf{X}))^2)$ car $\forall \mathbf{x} \in \mathbb{R}^d$,

$E((Y - g(\mathbf{X}))^2/\mathbf{X} = \mathbf{x})$

$= E((Y - E(Y/\mathbf{X}) + E(Y/\mathbf{X}) - g(\mathbf{X}))^2/\mathbf{X} = \mathbf{x})$

$= E((Y - E(Y/\mathbf{x}) + E(Y/\mathbf{x}) - g(\mathbf{x}))^2/\mathbf{X} = \mathbf{x})$

$= (E(Y/\mathbf{x}) - g(\mathbf{x}))^2 + 2(E(Y/\mathbf{x}) - g(\mathbf{x}))E(Y - E(Y/\mathbf{x})/\mathbf{X} = \mathbf{x})$

$$+E((Y - E(Y/\mathbf{x})/\mathbf{X} = \mathbf{x})^2)$$

$$= (E(Y/\mathbf{x}) - g(\mathbf{x}))^2 + E\left((Y - E(Y/\mathbf{x})/\mathbf{X} = \mathbf{x})^2\right). \quad (1.18)$$

En conséquence, la valeur minimale de $E((Y - g(\mathbf{X}))^2/\mathbf{X} = \mathbf{x})$ est donnée pour $g(\mathbf{x}) = E(Y/\mathbf{x})$, le minimum étant la variance conditionnelle de Y :

$E\left((Y - E(Y/\mathbf{x})/\mathbf{X} = \mathbf{x})^2\right)$.

Le minimum de l'erreur quadratique étant atteint pour chaque valeur de $\mathbf{x}$, il en résulte que fonction $g(.)$ qui minimise $E(Y - g(\mathbf{X}))^2$ est $g(\mathbf{X}) = E(Y/\mathbf{X})$. On constate que la décision à structure imposée est un cas particulier de la régression.

Le théorème suivant montre de plus que la probabilité d'erreur obtenue avec une règle de décision élaborée à partir d'un estimateur convergent de $p(\omega_1/\mathbf{x})$ tend asymptotiquement vers celle de la règle de Bayes.

Théorème 1.2.2 *Soit $\hat{p}_N(\omega_1/\mathbf{x})$ un estimateur faiblement convergent de $p(\omega_1/\mathbf{x})$, c'est à dire tel que :* $\lim_{N\to\infty} E\left((\hat{p}_N(\omega_1/\mathbf{x}) - p(\omega_1/\mathbf{x}))^2\right) = 0$, *et soit une règle de décision définie comme en [1.16], alors :*

$$\lim_{N\to\infty} \frac{P_e(d_N) - P_e^*}{\sqrt{E\left(\hat{p}_N(\omega_1/\mathbf{X}) - p(\omega_1/\mathbf{X}))^2\right)}} = 0. \quad (1.19)$$

Preuve : (voir par exemple [DEV 96])

Ce résultat montre clairement que $P_e(d_N) - P_e^*$ converge vers 0 plus vite que l'erreur quadratique de régression. On peut interpréter ce résultat en constatant qu'il suffit d'estimer correctement $p(\omega_1/\mathbf{x})$ pour les observations $\mathbf{x}$ pour lesquelles $p(\omega_1/\mathbf{x}) \approx 1/2$, c'est à dire uniquement au voisinage de la frontière entre les classes. En revanche, le théorème précédent ne donne pas d'indication quant à la vitesse de convergence de $\frac{P_e(d_N) - P_e^*}{\sqrt{E\left(\hat{p}_N(\omega_1/\mathbf{X}) - p(\omega_1/\mathbf{X}))^2\right)}}$ vers 0, qui peut être très lente (ou très rapide).

1.2.1. *Minimisation structurelle du risque et sélection de détecteur*

Nous allons évoquer maintenant le principe de Minimisation Structurelle du Risque (MSR). Supposons que nous disposions d'un ensemble de classes imbriquées ($C_k \subset C_{K+1}$) de familles de détecteurs et d'un intervalle de confiance (au niveau $1 - \alpha$) de la forme :

$$C(\theta) < C_e(\theta) + S(N, \alpha, V_c) \quad (1.20)$$

où S représente la largeur de l'intervalle de confiance, fonction de N, du niveau de confiance et de la "complexité" V_c de chaque famille (qui sera formellement définie plus loin). Pour chaque valeur de k, nous allons rechercher le détecteur qui minimise le risque empirique. Ainsi que nous le verrons, la borne supérieure de l'intervalle de confiance, que l'on appelle "risque garanti" est, à N et α fixés, une fonction de V_c. Elle est la somme du risque empirique, qui est une fonction décroissante de V_c (si la famille est suffisamment riche, il est possible d'apprendre par coeur l'ensemble d'apprentissage) et de la largeur de l'intervalle de confiance qui est une fonction croissante de V_c. Le risque garanti passe par un minimum. C'est la valeur de k qui correspond à ce minimum qui nous permet de retenir le meilleur détecteur.

En dehors des cas les plus simples, le calcul analytique du paramètre V_c est impossible. Aussi, afin d'appliquer le principe MSR, il faut avoir recours à des techniques numériques d'évaluation de V_c. Une alternative consiste à estimer le risque sur des données non utilisées pour l'apprentissage du détecteur à l'aide, par exemple, de techniques de validation croisée ou de ré-échantillonnage (voir section xxx). On retiendra alors le détecteur qui offre la meilleure estimation de ce risque. On parle alors de "minimisation structurelle du risque au sens large".

1.3. Compromis biais-variance

La section précédente a permis de faire le lien entre l'élaboration d'une règle de décision et la régression. Nous allons présenter dans ce paragraphe le compromis biais-variance qui joue un rôle central lors de la sélection d'une règle de décision. Reprenons l'équation [1.18]. Nous avons :

$$E((Y - g(\mathbf{X}))^2/\mathbf{X} = \mathbf{x}) = (E(Y/\mathbf{x}) - g(\mathbf{x}))^2 + E\left((Y - E(Y/\mathbf{x})/\mathbf{X} = \mathbf{x})^2\right).$$

Considérons le terme $(E(Y/\mathbf{x}) - g(\mathbf{x}))^2$, dans lequel la fonction $g(\mathbf{x})$ est remplacée par un estimateur de $p(\omega_1/\mathbf{x})$: $\hat{p}_N(\omega_1/\mathbf{x}) \equiv \hat{p}(\omega_1/\mathbf{x}; A_N)$. Le premier terme de droite de [1.18] devient aléatoire et il convient alors de prendre l'espérance par rapport à A_N. Nous avons alors :

$$E_{A_N}\left((E(Y/\mathbf{x}) - \hat{p}_N(\omega_1/\mathbf{x}))^2\right) = (E(Y/\mathbf{x}) - E_{A_N}(\hat{p}_N(\omega_1/\mathbf{x})))^2 + Var(\hat{p}_N(\omega_1/\mathbf{x})). \quad (1.21)$$

L'erreur quadratique définie en [1.18] est alors la somme de 3 termes :

- l'erreur minimale atteignable (indépendante du choix de $\hat{p}_N(\omega_1/\mathbf{x})$)
- un terme de biais (1^{er} terme de droite de [1.21])
- un terme de variance (2^{eme} terme de droite de [1.21]).

Considérons à nouveau un ensemble de classes C_k imbriquées ($C_k \subset C_{k+1}$) et, pour chaque valeur de k, recherchons le meilleur estimateur de $p(\omega_1/\mathbf{x})$. On observe en général que le terme dû au biais diminue avec k et, en revanche, que le terme de variance augmente avec k. L'erreur quadratique obtenue est donc une fonction de k, ainsi qu'il est montré sur la figure [1.1].

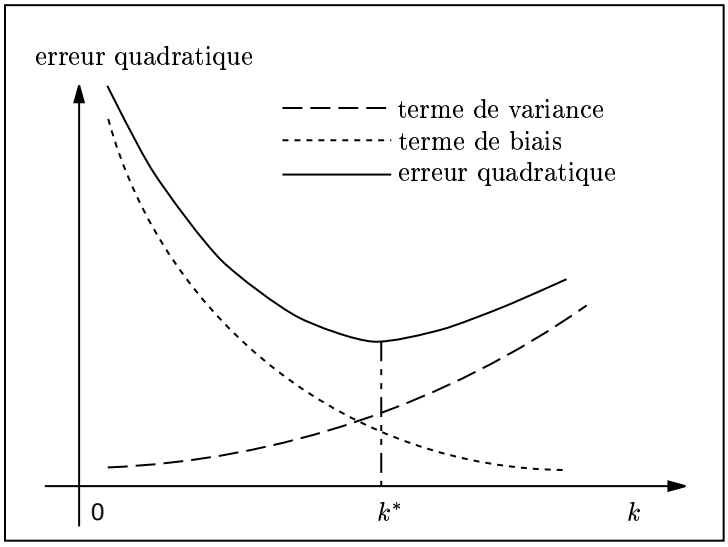

Figure 1.1 – Décomposition de l'erreur quadratique en fonction de k

Il est entendu que cette erreur quadratique ne peut être déterminée (la loi de probabilité des observations $p(\mathbf{x}, y)$ est inconnue). La recherche du "meilleur" estimateur $\hat{p}_N(\omega_1/\mathbf{x})$ pourra toutefois être effectuée par minimisation structurelle du risque au sens large, à l'aide de techniques de validation ou de ré échantillonnage (voir section xxx).

Une alternative consiste à considérer une classe C suffisamment "riche" et à restreindre la recherche au sein de celle-ci par des techniques de pénalisation ou de régularisation (voir section yyy).

Dans tous les cas, si $p(\omega_1/\mathbf{x}) \notin C_k \forall k$, on ne pourra espérer de bonnes performances asymptotiques du détecteur. En effet, la convergence de $\hat{p}_N(\omega_1/\mathbf{x})$ vers $p(\omega_1/\mathbf{x})$ ne pourra être assurée et la consistance de la règle de décision ne sera pas vérifiée.

1.4. Minimisation de l'erreur empirique

Dans l'introduction, nous avons vu que l'élaboration d'un détecteur reposait sur l'optimisation d'un critère de performance. Après avoir considéré l'erreur quadratique, nous allons maintenant étudier le cas où l'on retient l'erreur empirique, c'est à dire la proportion d'individus de l'ensemble d'apprentissage mal classés par la règle considérée. Nous allons voir que des considérations combinatoires permettent, en particulier, de caractériser les classes C_k évoquées précédemment et de proposer une méthode de sélection du "meilleur" détecteur.

La classe `hermes.cls` est prévue pour être utilisée en lieu et place de la classe `book.cls`. Pour l'activer on utilisera donc, après avoir placé le fichier `hermes.cls` à un endroit accessible à LaTeX 2_ε, la commande

```
\documentclass[10pt,options]{book}
```

La taille de fonte de base de 10pt correspond aux normes Hermes Science Publications. Les options disponibles sont décrites au point 1.5.

La classe `hermes` fait appel à la classe `book.cls` et aux extensions suivantes qui doivent donc être disponibles dans votre environnement LaTeX 2_ε :

- Extensions standard distribuées avec LaTeX2e
  - `fontenc`
  - `ifthen`
  - `babel`
- Extensions disponibles sur les archives TeX CTAN
  - `setspace`
  - `chicago`

L'extension `babel` est utilisée pour passer des règles de typographie de l'anglais aux règles du français, notamment pour ce qui est de l'espacement avant et après la ponctuation, les règles de césure et les intitulés (chapter - chapitre, table - tableau, etc.). Pour un fonctionnement correct des règles de césure, il faut impérativement utiliser un format `latex.fmt` comprenant les modèles de césures anglais et français. Pour créer ce format, il faut exécuter *initex* sur `latex.ltx` après avoir modifié le fichier `language.dat` comme suit

```
% File    : language.dat
% Purpose : specify which hyphenation patterns to load
%           while running iniTeX
english ushyphen.tex
french f8hyph.tex
```

1.5. Les options

La classe `hermes` accepte les options suivantes (les * indiquent les options de défaut) en plus de celles de la classe `book` :

`articles` Pour la production d'un ouvrage d'articles. Les chapitres ne sont pas numérotés. Le titre des chapitres (en fait le titre des articles) n'est pas précédé du mot « chapitre ». Il apparaît néanmoins dans la table des matières (ce qui n'est pas le cas avec `chapter*`). Il doit être introduit avec la commande `chapter[titrecourt]{titrelong}`. Si `titrecourt` est précisé, c'est celui-ci qui apparaît dans la table des matière.

`chapters(*)` Pour la production d'un ouvrage normal constitué de chapitres numérotés.

`hermesheadings(*)` Définit les titres courants selon les directives Hermes : La page de gauche (impaire) contient le titre de l'ouvrage qui doit être précisé avec la commande

```
\booktitle{titre de l'ouvrage}
```

Le titre courant de la page de droite doit être celui du chapitre et doit être précisé avec la commande `\markright{titrechapitre}`. Voir également ci-dessous les commandes `\includearticle`, `\includechapter`, `\newarticle` et `\newchapter`.

`myheadings` Si vous voulez utiliser des titres courants personnalisés. Equivalent à `\pagestyle{myheadings}`.

`headings` Equivalent à `\pagestyle{headings}`. Le titre du chapitre est utilisé comme titre courant sur les pages paires (gauche), et le titre de la section courante sur les pages impaires.

`freng(*)` Assure une homogénéité de présentation entre le français et l'anglais en particulier l'indentation après le titre des sections (qui change selon la langue choisie avec `babel`).

`french` Règles typographiques du français.

`english` Règles typographiques de l'anglais.

`bookbib` Les bibliographies débutent sur une nouvelle page, comme un chapitre. Valeur par défaut avec l'option `{chapters}`.

`articlebib` Les bibliographies débutent comme de nouvelles sections. Valeur par défaut avec l'option `{articles}`.

`chicaco` Active le style de citations bibliographiques `chicago`.

`chicagogr` Active une version modifiée de `chicago`, qui permet d'ajuster, avec `\begin{thebibliography}{texte}`, l'indentation à la longueur de `texte`.

`grbibstyle` Active un style de bibliographie simplifié qui permet d'ajuster, avec `\begin{thebibliography}{texte}`, l'indentation à la longueur de `texte` (ne fait pas appel à `chicago`.)

1.6. Les commandes

`\Fr` et `\Eng`
Ces commandes permettent d'activer le français (`\Fr`) ou l'anglais (`\Eng`). Elles sont à préférer au `\selectlanguage` de l'extension `babel`, car elles exécutent plusieurs modifications supplémentaires liées notamment aux commandes `\resume`, `\cles` et aux intitulés des figures et tableaux (voir ci-dessous).

```
\includearticle{auteurs}{titre}{nomfichier}{langue}
\includechapter{titre}{nomfichier}{langue}
\newarticle{auteurs}{titre}{langue}
\newchapter{titre}{langue}
```

où :

`auteurs` sont les noms des auteurs tels qu'ils apparaîtront dans la table des matières. Avec l'option `myheadings` c'est aussi le titre courant de défaut pour la page de gauche.

`titre` est le titre courant qui apparaît sur les pages impairs (en principe le titre du chapitre). Même effet que `\markright{titre}`.

`nomfichier` est le nom du fichier `.tex` du chapitre/article. L'extension `.tex` ne doit pas être donnée. Equivalent à `\include{nomfichier}`. On peut donc utiliser la commande `\includeonly{liste fichier}` pour la mise au point d'un sous-ensemble de chapitres.

`langue` soit `\Fr` pour un article en français et `\Eng` pour un article en anglais.

Les quatre commandes qui suivent assurent le cas échéant une présentation uniforme des noms d'auteurs, de leur affiliation, du résumé, des mots clés et des codes de classification. La présentation est celle que vous pouvez observer sur cet exemple. L'affiliation apparaît en particulier en note de pied.

```
\auteur{nom de l'auteur}
\affiliation{affiliation}
\resume{texte du résumé}
\cles{mots clés} \classif{codes de classification}
```

Enfin, les commandes suivantes permettent de gérer l'insertion de sections du type « annexe » dans un chapitre.

`\appendixinchap` modifie au format lettre capitale la numérotation des sections qui suivent. Utilisée avec l'option `articles`, cette commande produit un résultat équivalent à celui de `\appendix` avec la classe `article`. On peut également, lorsque `articles` est activé, utiliser `\appendix` qui est redéfini comme un alias de `\appendixinchap`.

`\stopappendix` redéfinit au format numérique la numérotation des sections. Le compteur `section` est remis à la valeur en vigueur lors du précédent `\appendixinchap`.

1.7. Les paramètres

Les paramètres `\bibspacing`, `\capspacing` et `\resuspacing`, définissent l'interlignage de la bibliographie pour le premier, des légendes des tableaux et figures pour le second, et du résumé et des mots clés pour le troisième. Ces valeurs peuvent être modifiées avec les commandes

```
\renewcommand*{\bibspacing}{.9}
\renewcommand*{\capspacing}{.8}
\renewcommand*{\resuspacing}{.8}
```

Les valeurs par défaut sont respectivement 0.9, 0.8 et 0.8.

Les paramètres `\frenchfigurename` et `\frenchtablename` permettent de contrôler le nom et le format de l'intitulé des figures et des tableaux qui sera employé en français. De même `\englishfigurename` et `\englishtablename` permettent de contrôler les intitulés anglais. Ces noms remplacent ceux fixés par l'extension `babel` qui sont « Figure » et « Table » lorsque l'anglais est sélectionné, et « Fig. » et « Tab. » lorsque le français est activé. La ponctuation et/ou l'espacement entre l'intitulé et la légende est donnée par `\caphsep`. On peut modifier les définitions utilisées par défaut avec les commandes

```
\renewcommand*{\englishfigurename}{Figure}
\renewcommand*{\englishtablename}{Table}}
\renewcommand*{\frenchfigurename}{Figure}
\renewcommand*{\frenchtablename}{Tableau}}
\renewcommand*{\caphsep}{.\hspace{1ex}}
```

Les exemples ci-dessus donnent les définitions utilisées par défaut.

Afin d'éviter les césures en fin de page, la valeur de `\brokenpenalty` est fixée à 4000 par `hermes.cls`.

environnement	français	anglais
table	Tableau	Table
figure	Figure	Figure

Tableau 1.1 – Intitulés générés par `\caption{...}`

Il faut éviter les lignes seules au début (widow) ou à la fin (club) d'une page. Les pénalités fixées par `hermes.cls` sont

```
\clubpenalty=3000
\widowpenalty=4000
```

Elles n'excluent pas la présence de veufs ou orphelins, mais forcent LaTeX à essayer fortement de les éviter. Pour interdire absolument les veufs et orphelins il faut fixer ces paramètres à 10000.

1.8. Bibliographie et références à la bibliographie

Ce manuel ainsi que l'exemple d'article qui l'accompagne illustrent la forme des références bibliographiques demandées par Hermes Science Publications.

Ce format de bibliographie peut être généré automatiquement avec le style `hermes.bst` à partir d'une ou de plusieurs bases de données bibliographiques standard BibTeX. Le style bibliographique `hermes` est une variante du style `alpha`. Les modifications apportées concernent

- les clés de références (trois premières lettres du premier auteur en capitales suivies d'une espace et des deux derniers chiffres de l'année de publication) ;
- le nom des auteurs en petites capitales ;
- la réduction de l'indentation après la première ligne d'un `bibitem` ;
- la possibilité de contrôler la langue du *and* avec la commande `\andname` (`\Fr` donne la valeur « et » à cette variable, et `\Eng` la fixe à « and ») ;
- nouveau champ `language` qui permet de préciserindividuellement avec `\Fr` et `\Eng` la langue de chaque entrée dans la base bibliographique.

Pour générer les bibliographies, il faut, pour une bibliographie unique à la fin de l'ouvrage, insérer les commandes suivantes à l'endroit où l'on veut placer la bibliographie.

```
\bibilographystyle{hermes}
\bibliography{liste des bases de données bibliographiques}
```

Après avoir compiler le fichier, il convient de « bibtex-er » le fichier principal, puis de recompiler. Par exemple, on exécutera la commande

```
bibtex hermes
```

qui génère le fichier `hermes.bbl` contenant la bibliographie qui sera insérée par `\bibliography{...}`.

Si l'on veut une bibliographie par chapitre, il faut insérer manuellement le fichier `.bbl` dans le fichier de chaque chapitre. Par exemple, pour le présent manuel, on a utilisé

```
\bibliographystyle{hermes}
\nocite{KopkaDa95,Rolland95,Gossen94,Eijkhout92}
\bibliography{herm}
\InputIfFileExists{herm_ch1.bbl}{}{}
```

où `herm_ch1` est le nom du source `.tex` de ce premier chapitre. La commande `\nocite` a été utilisée pour faire apparaître dans la bibliographie les références non citées dans le texte.

Il est évidemment possible de générer manuellement la bibliographie sans utiliser les bases de données bibliographiques et BibTeX. A titre d'exemple, voici le code qui génère la bibliographie de ce manuel.

```
\begin{thebibliography}{m}

\bibitem[EIJ~92]{Eijkhout92}
{\scshape V.~Eijkhout}.
\newblock  {\em \TeX\ by Topic: A \TeX{}nician's Reference}.
\newblock Addison-Wesley, Reading, Massachusetts, 1992.

\bibitem[GOO~94]{Gossen94}
{\scshape M.~Goossens, F.~Mittelbach \andname{} A.~Samarin}.
\newblock {\em The \LaTeX\ Companion}.
\newblock Addison-Wesley, Reading, Massachusetts, 1994.

\bibitem[KOP~95]{KopkaDa95}
{\scshape H.~Kopka \andname{} P.~W. Daly}.
\newblock {\em A Guide to \LaTeXe: Document Preparation for
Beginners and Advanced Users}.
\newblock Addison-Wesley, Reading, Massachusetts, 2nd
edition, 1995.

\bibitem[ROL~95]{Rolland95}
```

```
{\scshape C.~Rolland}.
\newblock {\em \LaTeX: guide pratique}.
\newblock Addison-Wesley, Paris, 2ème edition, 1995.

\end{thebibliography}
```

1.9. Quelques conseils

Hermes Science Publications suggère fortement d'éviter les césures en fin de page, spécialement en fin de page impaire. Le cas échéant, on peut éviter une césure non voulue en plaçant un `\linebreak[3]` ou un `\linebreak[4]` au début du mot coupé. La première proposition ne coupera la ligne à l'endroit indiqué que si cela peut se faire sans étirer exagérément le contenu de la ligne.

La table des matières est générée automatiquement par la commande `\tableofcontent` qui doit être insérée à l'endroit où cette table doit figurer. Le cas échéant, si le fichier à compiler s'appelle `exemple.tex`, on peut modifier le contenu de la table en éditant le fichier `exemple.toc`.

Avec l'option `articles`, seul le titre des articles apparaît dans la table des matières. Avec l'option `chapters` les sections de niveau 1 y figurent également. Le cas échéant la commande `\setcounter{tocdepth}{n}` permet de fixer le niveau n voulu. En LaTeX il existe 5 niveaux de sections dont la profondeur de numérotation peut être contrôlée avec `\setcounter{secnumdepth}{n}`. La valeur par défaut de `secnumdepth` est 2 avec l'option `articles` et 3 avec l'option `chapters`. Les commandes

```
\setcounter{secnumdepth}{5}
\section{Section}
\subsection{Subsection}
\subsubsection{Subsubsection} Texte texte ...
\paragraph{Paragraph} Texte texte.
\subparagraph{Subparagraph} Texte texte.
```

produisent les inter-titres illustrés au tableau 1.2. Avec l'option `chapters`, la numérotation est de plus précédée par le numéro du chapitre.

Les ouvrage Hermes Science devraient en principe contenir un index. Celui-ci peut facilement être généré en plaçant l'instruction `\makeindex` dans le préambule et en plaçant des `\index{mot à indexer}` aux endroits voulus dans le texte. Ne pas oublier de générer le fichier `.ind` avec l'utilitaire `makeindex.exe`. Par exemple, si le fichier principal est `exemple.tex`, il faut exécuter, après avoir compiler le fichier `exemple.tex`, la commande

1.1. Section

1.1.1. *Subsection*

1.1.1.1. *Subsubsection*

Texte texte ...

1.1.1.1.1. Paragraph Texte texte ...

1.1.1.1.1.1. *Subparagraph* Texte texte ...

Tableau 1.2 – Inter-titres avec `hermes.cls`

```
makeindex exemple
```

L'index est ensuite inséré dans le document en mettant à l'endroit désiré `\input{exemple.ind}`. Utiliser `\addcontentsline{toc}{chapter}{Index}` pour faire apparaître l'index dans la table des matières.

Les manuscrits soumis aux éditions Hermès doivent commencer à la page 5. Ceci se fait en insérant dans le préambule `\setcounter{page}{5}`.

Bibliographie

[BEL 61] R. Bellman. *Adaptive Control Processes : a Guided Tour.* New Jersey : Princeton University Press, 1961.

[DEV 96] L. Devroye, L. Györfi, G. Lugosi. *A probabilistic Theory of Pattern Recognition.* Applications of Mathematics : Springer, 1996.

[FUK 72] K. Fukunaga. *An Introduction to statistical pattern recognition.* Addison-Wesley (2ème édition 1990), 1972.

[NAR 77] P. Narendra and K. Fukunaga. *A branch and bound algorithm for feature subset selection.* IEEE transactions on Computers, 26 : 917-922.

[STO 77] C. Stone. *Consistent non parametric regression.* Annals of statistics, 8 : 1348-1360, 1977.

[EIJ 92] V. Eijkhout. *TEX by Topic : A TEXnician's Reference.* Addison-Wesley, Reading, Massachusetts, 1992.

[GOO 94] M. Goossens, F. Mittelbach et A. Samarin. *The LATEX Companion.* Addison-Wesley, Reading, Massachusetts, 1994.

[KOP 95] H. Kopka et P. W. Daly. *A Guide to LATEX 2ε : Document Preparation for Beginners and Advanced Users.* Addison-Wesley, Reading, Massachusetts, 2nd edition, 1995.

[ROL 95] C. Rolland. *LATEX : guide pratique.* Addison-Wesley, Paris, 2ème edition, 1995.