5. LA SÉLECTION DE MODÈLES

Résumé

Pour résoudre un problème de modélisation ou de classification, deux étapes se succèdent généralement dans la conception du modèle ou du classifieur :

- *En premier lieu, il faut choisir des variables descriptives (ou descripteurs, ou facteurs) pertinentes, c'est-à-dire les variables qui agissent sur la sortie du processus à modéliser (dans le cas de la modélisation de processus), ou qui déterminent la classe de l'objet à classer (dans le cas d'un problème de classification) ; on suppose qu'il existe une relation entre ces variables et la sortie ou la classe désirée.*
- *En second lieu, on cherche, dans une famille de fonctions, celle qui permet d'estimer "au mieux" la valeur de la sortie mesurée du phénomène à partir des valeurs des descripteurs.*

Ainsi, dans une procédure de modélisation, on dispose d'un ensemble de descripteurs à partir desquels il est possible de construire un ensemble de modèles. L'objectif des méthodes de sélection de modèles est de choisir, parmi cet ensemble de modèles, celui qui explique le mieux, les phénomènes observés. Plus précisément, comme nous l'avons souligné plus haut, on cherche le modèle le plus simple qui atteigne les performances spécifiées dans le cahier des charges.

La première partie de ce chapitre présente les bases des méthodes de sélection de modèles les plus fréquemment utilisées. Nous décrivons ensuite des méthodes partielles de sélection de modèles, plus économes en temps de calcul.

Dans une deuxième partie, nous présentons une procédure originale de sélection de modèles ; enfin, nous présentons une application de cette méthode à la sélection de l'architecture d'un réseau de neurones à une couche cachée. Cette méthode originale nous permet ainsi de franchir, de façon presque automatique, les deux étapes de la modélisation indiquées ci-dessus.

Toutes les méthodes présentées dans ce chapitre ont une justification théorique dans le cadre de la modélisation de processus. Nous montrerons qu'elles peuvent également s'appliquer aux problèmes de classification, mais qu'il faut être conscient de leurs limitations lorsqu'on les met en œuvre dans ce cadre.

5.1 Introduction

Pour résoudre un problème de modélisation [voir par exemple Urbani 95], il faut effectuer plusieurs choix :

- Choix du ***type*** du modèle ; c'est-à-dire des caractéristiques générales du modèle (par exemple modèle statique ou dynamique, modèle linéaire ou non linéaire, ...). De

façon pratique, le choix du type du modèle est résolu par l'analyse de la nature des phénomènes observés. Ainsi, un problème de classification sera, dans la plupart des cas, considéré comme un problème statique ; en revanche, la modélisation de processus met généralement en jeu des modèles dynamiques.

- Choix, si c'est possible, d'un ***ensemble de points expérimentaux*** (plan d'expérience) fournissant l'ensemble d'apprentissage.

  Ces deux points ne seront pas abordés dans ce travail car ils relèvent du domaine de compétence de l'ingénieur. Nous supposerons seulement qu'une étude préalable a conduit à choisir un type de modèle et un ensemble de points expérimentaux.

- Choix de la ***structure*** du modèle, c'est-à-dire celui d'une famille F de fonctions (par exemple modèle linéaire, réseau de neurones, réseau d'ondelettes) et par l'ensemble des variables descriptives nécessaires.

On peut alors procéder à l'estimation des ***paramètres*** $\boldsymbol{\theta}$ du modèle, qui déterminent la fonction choisie au sein de la famille F. Cette étape de la construction du modèle a été décrite dans le chapitre précédent.

Dans le présent chapitre, nous abordons la sélection de la structure du modèle. La structure d'un modèle parmi un ensemble de candidats se fait en comparant les performances du meilleur modèle de chaque structure, après l'estimation des paramètres.

Dans une première partie, nous nous intéressons à la première étape de la détermination de la structure : le choix des descripteurs pertinents. Nous nous placerons dans le cadre de la mise en œuvre de modèles linéaires par rapport aux paramètres. Nous décrirons :

- les outils de comparaisons entre deux modèles,
- la procédure optimale de sélection de modèles,
- deux procédures qui permettent de réduire considérablement le nombre de modèles à évaluer.

Enfin, la dernière partie du chapitre propose une méthode originale qui présente deux avantages : d'une part, elle est économe en nombre de modèles testés pour la sélection, et, d'autre part, elle est applicable à la sélection de l'architecture d'un réseau de neurones à une couche cachée. Nous pouvons ainsi traiter la deuxième étape du choix de la structure : le choix de la famille de fonctions susceptibles de modéliser le phénomène.

5.2 Comparaison entre modèles

Comme nous l'avons indiqué plus haut, la sélection de structures s'effectue en comparant les performances des modèles candidats. Bien entendu, il n'est pas possible d'évaluer les performances d'un modèle en testant de manière exhaustive son comportement dans toutes les situations qu'il est susceptible de modéliser : on ne peut effectuer ce test que sur un nombre *fini* d'échantillons. La comparaison de performances d'un modèle présente donc un caractère statistique. Les *tests d'hypothèses* sont les outils qui sont généralement utilisés

pour résoudre ces problèmes de comparaison de performances [voir par exemple Grais 92]. Nous commençons donc par une brève présentation des tests d'hypothèses.

Un exemple classique d'utilisation des tests d'hypothèses est le contrôle de pièces de fabrication. Le problème est le suivant : après avoir estimé la dimension moyenne d'un échantillon de pièces prélevé sur une ligne de fabrication, on cherche à savoir si cette estimation de la dimension est, ou n'est pas, significativement différente de celle spécifiée dans le cahier des charges.

5.2.1 Principe

Supposons qu'au cours de la fabrication, on estime le diamètre moyen (noté δ) d'un échantillon de pièces. Dans le cas général, cette estimation est différente du diamètre spécifié (noté δ_0). Avant de se lancer dans un réglage long et coûteux des machines suspectées, il faut connaître les causes possibles de l'écart observé entre δ et δ_0. En effet, il peut avoir deux origines :

- il est dû aux fluctuations aléatoires,
- il est effectivement dû à un dérèglement ou à l'usure de la machine.

Il s'agit de choisir entre ces deux hypothèses et de décider si l'écart observé est significatif (avec un seuil de risque d'erreur fixé) et rend compte d'une différence réelle ou, au contraire, n'est pas significatif et est dû au hasard.

5.2.2 Description

On définit deux hypothèses alternatives H_0 et H_1 que l'on désire tester :

- H_0 : $\delta = \delta_0$ (hypothèse nulle),
- H_1 : $\delta \neq \delta_0$ (hypothèse alternative).

La procédure habituelle de test d'hypothèses est la suivante : on considère H_0 comme exacte ; dans cette hypothèse, l'écart observé ne peut être dû qu'aux seules fluctuations résultant de l'échantillonnage. On en déduit alors la loi de distribution de la proportion d'erreurs.

On se fixe une probabilité α ou risque d'erreur que l'on juge acceptable. La probabilité α correspond au risque acceptable de rejeter H_0 alors qu'elle est vraie (et donc d'adopter H_1) :

$$\alpha = P\{rejeter\ H_0 / H_0\ vraie\}$$: risque de 1ère espèce

Par exemple, en prenant $\alpha = 0.05$, on accepte 5 chances sur 100 de considérer que le lot de pièces présente une espérance mathématique (du diamètre) différente de δ_0 alors que, en réalité, celle-ci est égale à δ_0.

La probabilité α définit donc la région d'acceptation de l'écart observé. Ainsi,

- si le diamètre observé δ observé n'appartient pas à la région d'acceptation, on rejette H_0 et l'on retient H_1,

- si le diamètre observé δ observé appartient à la région d'acceptation, alors rien ne s'oppose à ce que l'on accepte H_0 (les données dont on dispose ne sont pas en contradiction avec cette hypothèse).

5.2.3 Le test de Fisher

Le test de Fisher est le test d'hypothèses le plus utilisé dans le cas de modèles linéaires par rapport aux paramètres. Il est applicable lorsque l'on cherche à comparer un modèle complet à un sous-modèle plus restreint.

5.2.3.1 Principe

Pour la sélection de modèle, nous supposons que le modèle complet obéit à l'équation suivante (les vecteurs des entrées et de la sortie sont centrés) :

$$Y = X\,\theta_P + \omega$$

avec Y : vecteur aléatoire de dimension N (N est le nombre d'exemples),

θ_P : vecteur de dimension P des paramètres inconnus du modèle (P est le nombre de descripteurs),

X : matrice des entrées, de dimension $N \times P$ (P colonnes correspondant aux P descripteurs du modèle, et N lignes représentant les N exemples),

ω : vecteur du bruit, centré, non corrélé, de dimension N, normalement distribué (de moyenne nulle et de variance σ^2).

Ces hypothèses impliquent que le modèle défini ci-dessus est *complet,* c'est-à-dire qu'il contient la fonction de régression.

Tester si l'effet d'un ou plusieurs descripteurs parmi les P initiaux est statistiquement significatif revient à tester l'hypothèse de nullité des q coefficients correspondants.

On définit :

$Y_{mc}^{(complet)} = X\,\theta_{mc}^{(complet)}$: solution des moindres carrés,

$Y_{mc}^{(incomplet)} = X\,\theta_{mc}^{(incomplet)}$: solution des moindres carrés sous la contrainte des q coefficients nuls,

et la variable aléatoire $T^2 = \dfrac{N-P-1}{q} \cdot \dfrac{\left\|Y - Y_{mc}^{(incomplet)}\right\|^2 - \left\|Y - Y_{mc}^{(complet)}\right\|^2}{\left\|Y - Y_{mc}^{(complet)}\right\|^2}$

Pour ce test, les deux hypothèses alternatives sont :

- H_0 : les q coefficients sont nuls
- H_1 : les q coefficients ne sont pas nuls

Si H_0 est vraie (hypothèse nulle), alors la variable aléatoire T^2 suit une loi de Fisher-Snedecor à q et $(N\text{-}P\text{-}1)$ degrés de liberté ; ce qui permet de tester H_0 à partir de la valeur de la réalisation de T^2 dont on dispose. Si le test conduit à rejeter H_0 alors le sous-modèle est rejeté.

Il faut souligner que les tests d'hypothèses statistiques comparent un sous-modèle au modèle complet ; il est donc nécessaire d'avoir une relation d'inclusion entre les deux. D'autres tests d'hypothèses, qui ne nécessitent pas cette relation, ont été proposés, tels que les tests TRV (Test du Rapport de Vraisemblance) [Goodwin 77] et LDRT (Logarithm Determinant Ratio Test) [Leontaritis 87]. [Söderström 77] montre que ces tests et le test de Fisher sont asymptotiquement équivalents.

5.2.3.2 Mise en œuvre

Le principe de l'utilisation du test de Fisher (ou d'autres tests d'hypothèses statistiques) pour la sélection de modèles est d'accepter ou de rejeter un sous-modèle par rapport au modèle complet. En pratique, on part du modèle complet, avec tous les descripteurs dont on dispose, et l'on construit tous les sous-modèles possibles (on peut se limiter à un sous-ensemble des sous-modèles, comme nous le verrons plus loin). Ensuite, on compare, à l'aide du test de Fisher, le modèle complet à chacun des sous-modèles. Plusieurs voies sont alors possibles :

- si tous les sous-modèles sont rejetés, le modèle sélectionné est le modèle complet,
- si un ou plusieurs sous-modèles ne sont pas rejetés, il faut en choisir un. Comme il n'y a pas de relation d'inclusion entre eux on ne peut plus les comparer à l'aide du test. On choisit alors le modèle le moins complexe (celui qui possède le plus petit nombre de paramètres à ajuster par exemple) ; si, là encore, plusieurs sous-modèles possèdent la même complexité, on choisit celui qui minimise l'écart quadratique moyen.

Cette procédure est simple d'utilisation, mais elle nécessite un nombre très vite prohibitif d'estimations de paramètres. En effet, le nombre de sous-modèles à considérer à partir d'un modèle complet à P descripteurs est de 2^P. Le paragraphe 5.3 présente des méthodes permettant de réduire ce nombre.

5.2.4 Critère d'Information d'Akaike (AIC)

Les méthodes décrites au paragraphe précédent sont fondés sur la comparaison des performances exprimées par l'erreur quadratique sur un ensemble d'échantillons. Une autre approche consiste à construire une fonction de coût (ou indice de performance) qui tienne compte à la fois de la performance du modèle et de sa complexité. Le modèle conduisant à la plus petite valeur du critère est sélectionné. Ainsi, l'indice de performance peut être calculé pour deux modèles indépendants (sans qu'il soit nécessaire que l'un soit un sous-modèle de l'autre), et peut donc les comparer. Nous présentons ici le critère d'Akaike, qui est défini et discuté dans [Akaike 74], [Fourdrinier 94], [Urbani 95], [Chen 89], [Norton 86].

5.2.4.1 Définition

Le critère d'information d'Akaike est défini par :

$$AIC(\phi) = N \cdot \log(EQM) + P \cdot \phi$$

avec N : nombre d'exemples,

P : nombre de paramètres du modèle (correspond au nombre de descripteurs pour un modèle linéaire par rapport aux paramètres),

EQM : erreur quadratique moyenne des résidus (moyenne des carrés des écarts entre le modèle et les observations),

ϕ : facteur correspondant à la valeur de la distribution du χ^2 à un degré de liberté pour un niveau de confiance donné.

Dans la formulation du critère on reconnaît deux termes :

- le premier terme correspond à la performance du modèle : plus la performance est grande, plus l'écart entre la sortie du modèle et la sortie mesurée est faible, donc plus son logarithme est petit.
- le deuxième terme exprime la complexité du modèle, qui est proportionnelle au nombre de paramètres de celui-ci.

Pour utiliser ce critère, il faut spécifier un niveau de confiance et, par conséquent, choisir une valeur numérique pour ϕ. La valeur $\phi = 2$ a été critiquée car elle conduit, en général, à sélectionner des modèles plutôt sur-dimensionnés [Shibata 76]. [Chen 89] propose $\phi = 4$ qui est une valeur plus judicieuse correspondant à un risque de 1ère espèce calculé dans le cas où un modèle est un sous-modèle de l'autre environ égal à 0.05.

[Söderström 77] et [Leontaritis 87] ont étudié les relations entre le critère d'information d'Akaike et les autres tests statistiques.

5.3.4.2 Mise en œuvre

La mise en œuvre de la sélection de modèles avec un critère comme le critère d'information d'Akaike (AIC(4) par exemple) est semblable à celle des tests statistiques d'hypothèses : on se donne un ensemble de modèles à évaluer ; pour chacun d'entre eux, on calcule la valeur du critère et l'on choisit le modèle qui le minimise. L'avantage du critère d'information d'Akaike est qu'il permet de comparer deux modèles indépendants.

5.3 Stratégies de sélection de modèles

Comme nous l'avons indiqué plus haut, la méthode la plus "naturelle" consiste à évaluer tous les sous-modèles d'un modèle complet, ce qui peut exiger des temps de calcul très importants. D'autres stratégies, plus économes, réduisent le nombre de sous-modèles à estimer. Nous discutons dans ce paragraphe ces aspects de la mise en œuvre des tests d'hypothèses.

5.3.1 Stratégie exhaustive

Nous avons vu que la première approche possible pour choisir un modèle est de considérer un modèle complet, et de fabriquer tous les sous-modèles possibles, pour ensuite choisir le meilleur. Avec cette méthode, nous pouvons engendrer 2^P modèles à partir d'un ensemble de P descripteurs. Il faut donc estimer les paramètres de ces 2^P modèles et calculer les valeurs du test ou du critère qui leur sont associées. Le nombre de modèles à évaluer croît donc exponentiellement avec le nombre de descripteurs : la procédure devient rapidement impraticable. Elle reste néanmoins la méthode "optimale" de sélection puisque tous les sous-modèles sont évalués.

Nous décrivons deux méthodes "partielles", beaucoup plus économes en terme de nombre de modèles à tester, et qui, bien que sous-optimales en principe, possèdent toutefois de bonnes chances de mener au modèle optimal.

5.3.2 Stratégie "destructive"

L'idée de cette méthode, également nommée SBE (Stepwise Backward Elimination), est de considérer un modèle complet et d'en éliminer le descripteur le moins significatif. Autrement dit, on part du modèle complet à P descripteurs, on construit tous les sous-modèles possibles à P-1 descripteurs (soit P sous-modèles). On choisit celui qui offre la meilleure performance. Ensuite, on calcule les valeurs d'un critère de comparaison du sous-modèle choisi et du modèle complet :

- si le sous-modèle est meilleur (au sens du critère retenu) que le modèle complet on reprend la procédure à partir de ce sous-modèle (qui devient alors le modèle complet),
- sinon, on arrête la procédure et l'on conserve le modèle complet.

Le nombre maximal de modèles à considérer est :

$$1+\frac{P\cdot\left(P+1\right)}{2}$$

Ainsi pour traiter le problème donné ci-dessus à 15 descripteurs, il faudra construire, au plus, 121 sous-modèles.

5.3.3 Stratégie "constructive"

C'est la méthode symétrique de la méthode "destructive". Le point de départ est un modèle à 0 descripteur (seulement un terme constant : description par la moyenne des mesures) ; on construit les P sur-modèles à 1 descripteur. On choisit le meilleur modèle au sens du critère et l'on poursuit la procédure. On l'arrête lorsque le modèle est meilleur que tous ses sur-modèles.

Là encore, le nombre maximal de modèles à considérer est :

$$1+\frac{P\cdot(P+1)}{2}$$

5.3.4 En résumé

Par rapport à la stratégie exhaustive de sélection de modèles, les stratégies "partielles" ne conduisent pas toujours au modèle "optimal". Il ne faut pas pour autant les rejeter, car la méthode exhaustive dépasse très rapidement les capacités de calcul des calculateurs actuels. On est donc fréquemment contraint d'utiliser les méthodes partielles de sélection. Il est néanmoins possible de se faire une idée de la valeur du modèle sélectionné. En effet, lorsque les deux méthodes "constructive" et "destructive" conduisent au même modèle, on peut penser que celui-ci donnera des résultats très satisfaisants.

Comme nous l'avons souligné plus haut, les stratégies que nous venons de présenter ont l'inconvénient de nécessiter de nombreuses estimations de paramètres. Par exemple, dans le cadre d'une méthode "destructive" à P descripteurs initiaux il faut, pour choisir le meilleur sous-modèle à P-1 descripteurs, estimer les paramètres du modèle complet à P descripteurs et les paramètres des P sous-modèles à P-1 descripteurs.

Dans le paragraphe suivant, nous allons proposer une méthode originale de sélection de modèles linéaires par rapport aux paramètres, qui permet de pallier cet inconvénient.

5.4 Une méthode originale de sélection de modèles

Cette méthode s'applique à la modélisation d'un processus avec un modèle linéaire par rapport à ses paramètres.

5.4.1 Principe

L'idée de cette méthode est, dans un premier temps, d'ordonner les descripteurs par ordre d'importance, comme pour une approche "constructive". Au départ, nous disposons d'un ensemble de P descripteurs que l'on suppose suffisamment grand pour décrire les données (modèle complet). Parmi les P descripteurs, on cherche, à l'aide d'une méthode qui sera décrite dans le paragraphe 5.4.2, celui qui décrit le mieux la sortie désirée du processus, puis le deuxième et ainsi de suite. On obtient finalement un classement des descripteurs. Ensuite il faut considérer les P sous-modèles suivants :

- le 1er sous-modèle met en œuvre le 1er descripteur,
- le 2ème sous-modèle met en œuvre les 2 premiers descripteurs,
- le 3ème sous-modèle met en œuvre les 3 premiers descripteurs,
- ...
- le P-ième sous-modèle met en œuvre l'ensemble des P descripteurs[1].

[1] On peut également prendre pour 1er sous-modèle celui qui comprend uniquement un terme constant (modèle à 0 descripteur). Dans ce cas, le nombre total de sous-modèles à traiter est égal à P+1.

Le paragraphe 5.4.3 décrit la méthode qui nous permet de sélectionner le meilleur sous-modèle.

Cette méthode permet de considérer un nombre très réduit de modèles, et elle présente l'intérêt de bien faire prendre conscience de la pertinence relative de certains descripteurs pour le problème posé[2].

5.4.2 Classement des descripteurs

Le classement des descripteurs constitue la première étape de la procédure de sélection de modèles proposée. Il repose sur l'utilisation de l'algorithme d'orthogonalisation de Gram-Schmidt modifié. [Chen 89] en donne une description très détaillée. Nous ne reprendrons ici que le principe de l'algorithme ainsi qu'une interprétation géométrique [Urbani 95].

5.4.2.1 Algorithme de Gram-Schmidt modifié

Il existe deux manières de mettre en œuvre l'algorithme de Gram-Schmidt ; la première est dite classique (CGS : Classical Gram-Schmidt) et la seconde modifiée (MGS : Modified Gram-Schmidt). CGS est plus économe en terme d'occupation de la mémoire mais elle est très sensible aux erreurs d'arrondi [Björck 67]. La méthode MGS est numériquement plus stable. Il faut noter que ces deux méthodes seraient strictement équivalentes en l'absence d'erreurs d'arrondi. Comme la taille de la mémoire des machines à notre disposition le permet, nous utilisons l'algorithme de Gram-Schmidt modifié (MGS).

L'algorithme d'orthogonalisation de Gram-Schmidt considère les valeurs prises par les descripteurs et la sortie désirée comme des vecteurs. Les notations sont les suivantes :

$$X = \begin{bmatrix} x_1^1 & \cdots & x_P^1 \\ \vdots & & \vdots \\ x_1^N & \cdots & x_P^N \end{bmatrix} = \begin{bmatrix} X_1 & \cdots & X_P \end{bmatrix} \text{ matrice des entrées,}$$

$$\text{avec } X_p = \begin{bmatrix} x_p^1 \\ \vdots \\ x_p^N \end{bmatrix} \text{ vecteur de l'entrée } p,$$

$$Y = \begin{bmatrix} y^1 \\ \vdots \\ y^N \end{bmatrix} \text{ vecteur de la sortie.}$$

La matrice X est la matrice des entrées (P colonnes correspondant aux P descripteurs du modèle et N lignes représentant les N exemples de l'ensemble d'apprentissage). La matrice X est composée de P vecteurs représentant chacun une entrée. Le vecteur Y est le vecteur de

---

[2] Ce point est important car, de façon pratique, l'utilisateur a souvent fait beaucoup d'efforts pour obtenir ces descripteurs et il n'est pas toujours convaincu que certains d'entre eux doivent être éliminés.

sortie (N sorties observées des N exemples). Les vecteurs des entrées (X_p) et de la sortie (Y) sont centrés.

A la première itération, il faut trouver le vecteur d'entrée qui "explique" le mieux la sortie. Pour cela, on calcule le carré des cosinus des angles entre le vecteur de sortie et les vecteurs d'entrée :

$$\cos^2\left(X_p, Y\right) = \frac{\left(X_p^T\, Y\right)^2}{\left(X_p^T\, X_p\right)\cdot\left(Y^T\, Y\right)}$$

Le vecteur sélectionné est celui pour lequel cette quantité est maximale. Ensuite, on élimine la contribution de l'entrée sélectionnée en projetant le vecteur de sortie, et tous les vecteurs d'entrée restants, sur le sous-espace orthogonal au vecteur sélectionné.

La procédure se poursuit en choisissant, une nouvelle fois, le vecteur d'entrée projeté qui explique le mieux la sortie projetée. Elle se termine lorsque tous les vecteurs d'entrée ont été ordonnés.

Il faut souligner que l'estimation des moindres carrés ordinaires des paramètres s'obtient par la résolution immédiate d'une équation linéaire dont la matrice est triangulaire supérieure ; la norme du vecteur de sortie projeté détermine la valeur de l'EQM [Chen 89].

5.4.2.2 Interprétation géométrique

La figure 5.1 donne une interprétation géométrique de l'algorithme de Gram-Schmidt.

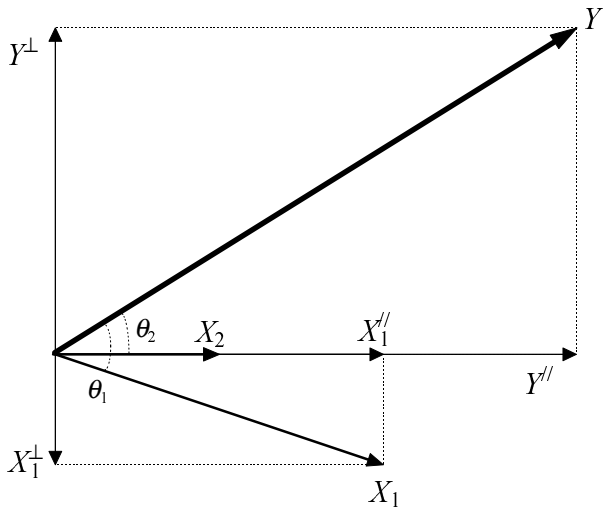


Figure 5.1 : Interprétation géométrique

Sur la figure, l'espace est de dimension 2. Le vecteur de sortie Y est mieux expliqué par le vecteur X_2 que par X_1 (l'angle θ_2 est plus petit que θ_1). De ce fait, X_2 est sélectionné par la méthode comme premier descripteur. Pour éliminer la partie expliquée par ce descripteur, on projette Y et X_1 (et plus généralement tous les vecteurs restants) sur le sous-espace orthogonal à X_2 ; on note ces projections $Y^{\perp}$ et $X_1^{\perp}$. Ici, on ne peut plus continuer et le dernier descripteur sélectionné est X_1.

L'écart quadratique moyen obtenu par les moindres carrés ordinaires avec le modèle à 1 descripteur (ici X_2) est donné par le carré de la norme du vecteur de sortie projeté (ici $Y^{\perp}$) divisé par le nombre d'exemples (ici 2 exemples).

5.4.2.3 "Sous-optimalité" de la procédure de Gram-Schmidt

Le principal intérêt de cette procédure est de ne considérer que P sous-modèles à partir des P descripteurs initiaux. De plus l'algorithme de Gram-Schmidt effectue simultanément, à chaque itération, deux opérations qui sont d'une part, le choix du meilleur descripteur, et, d'autre part, l'estimation des moindres carrés. Comme nous l'avons indiqué précédemment, cette procédure n'est pas "optimale". Le problème fictif suivant permettra de constater que cette procédure ne s'éloigne cependant pas beaucoup de la limite "optimale" [Lagarde 83]. Il sera repris pour illustrer la méthode sélection de modèles que nous proposons plus loin.

Considérons donc le problème comportant 15 points d'apprentissage et 10 descripteurs (dont 5 seulement sont pertinents) :

$$y^i = \sum_{p=1}^{10} \theta_p \, x_p^i + \omega^i \quad (i \text{ variant de 1 à 15})$$

avec x_p^i : $p^{\text{ème}}$ entrée distribuée suivant une loi de Gauss centrée et réduite,

θ_p : paramètres de la simulation (distribués suivant une loi de Gauss centrée et réduite pour p = 1, 2, 3, 4, 5 et nuls pour p = 6, 7, 8, 9, 10),

ω^i : bruit gaussien de variance égale à 2.10^{-2} (de moyenne nulle).

On note également :

$$\sigma^2 = \frac{1}{15} \cdot \sum_{i=1}^{15} (\omega^i)^2$$

La figure 5.2 compare la performance des modèles obtenus avec la procédure de Gram-Schmidt à celle de l'ensemble complet des 1024 sous-modèles.

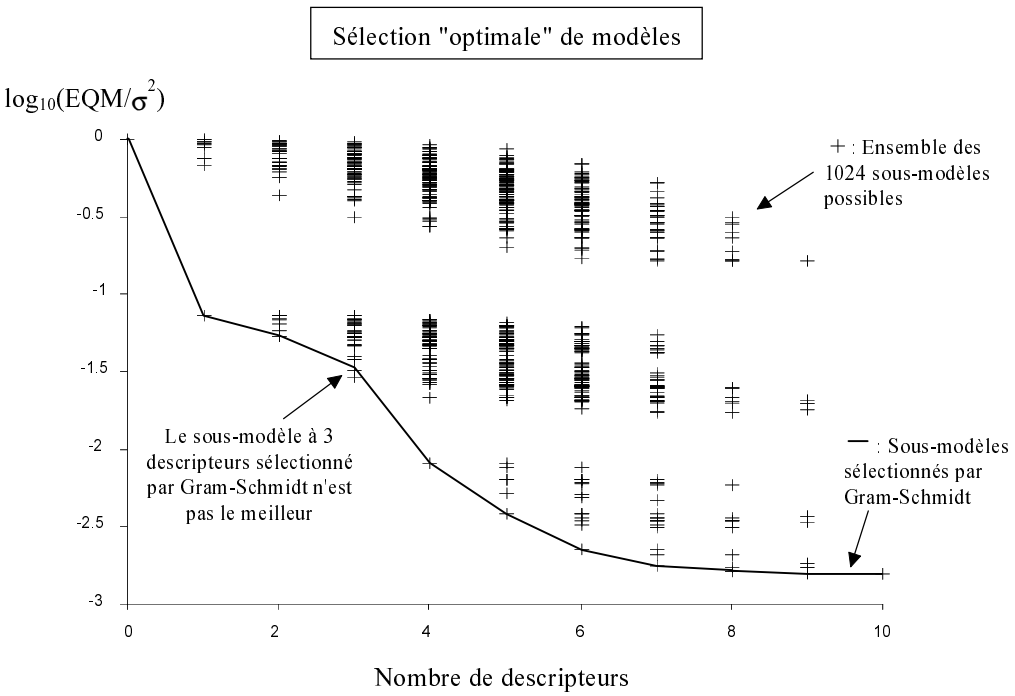


Figure 5.2 : Sélection "optimale" de modèles

Sur la figure, nous constatons que, pour les sous-modèles à 3 descripteurs, celui qui est choisi par la procédure de Gram-Schmidt n'est pas le meilleur. En effet, parmi les $C_{10}^{3} = 120$ sous-modèles possibles à 3 descripteurs, il en existe un qui offre une meilleure performance que celui sélectionné par Gram-Schmidt. Toutefois, la figure montre également que celui-ci n'est certes pas "optimal" mais qu'il n'est pas très éloigné de l'optimum.

On remarque également, sur la figure 5.2, qu'à partir d'un certain nombre de descripteurs, le gain en performance devient négligeable (l'EQM ne diminue plus beaucoup à partir de 7 descripteurs) ; en revanche, la complexité augmente. Il faut donc trouver un moyen de choisir un sous-modèle parmi les P sous-modèles donnés par la procédure de Gram-Schmidt. C'est l'objet du paragraphe suivant.

5.4.3 Choix du modèle

La sélection du modèle parmi les P initiaux peut être confiée à un test d'hypothèses statistiques ou à un critère comme le critère d'information d'Akaike. Nous proposons ici une méthode originale, fondée sur l'ajout d'un descripteur aléatoire.

L'idée est d'utiliser dans le modèle, outre les P descripteurs initiaux, un "descripteur aléatoire". Ensuite on utilise l'algorithme d'orthogonalisation de Gram-Schmidt décrit précédemment pour ordonner les P+1 descripteurs ainsi définis (le descripteur aléatoire et les P descripteurs initiaux). Les descripteurs rangés après la variable aléatoire sont considérés comme non pertinents pour le problème posé.

De façon pratique, on ordonne de la sorte une centaine de réalisations de la variable aléatoire pour obtenir la répartition du classement de la variable aléatoire.

Remarque : Du point de vue de l'organisation des calculs, il est plus intéressant de créer un ensemble de réalisations du descripteur aléatoire, et de lancer l'orthogonalisation de Gram-Schmidt sur la totalité des entrées (descripteurs initiaux + réalisations du descripteur aléatoire). A chaque itération, on choisit le meilleur descripteur en ne tenant pas compte des variables aléatoires. Une fois celui-ci choisi, on détecte les variables aléatoires qui expliquent mieux la sortie, on les compte et on les extrait du lot. Il ne reste plus qu'à projeter la sortie, les descripteurs restants et les variables aléatoires restantes sur le sous-espace orthogonal au descripteur sélectionné. A l'itération suivante, on procède de la même façon avec les descripteurs et les variables aléatoires restants. Ainsi, on n'effectue qu'une seule fois la procédure de Gram-Schmidt.

La figure 5.3 reprend le problème du paragraphe précédent, et montre :

- les valeurs de l'EQM prises par les sous-modèles sélectionnés par Gram-Schmidt (échelle de gauche, courbe décroissante),
- et le diagramme des fréquences cumulées (estimation de la fonction de répartition) du classement de 100 réalisations du descripteur aléatoire (échelle de droite, courbe croissante).

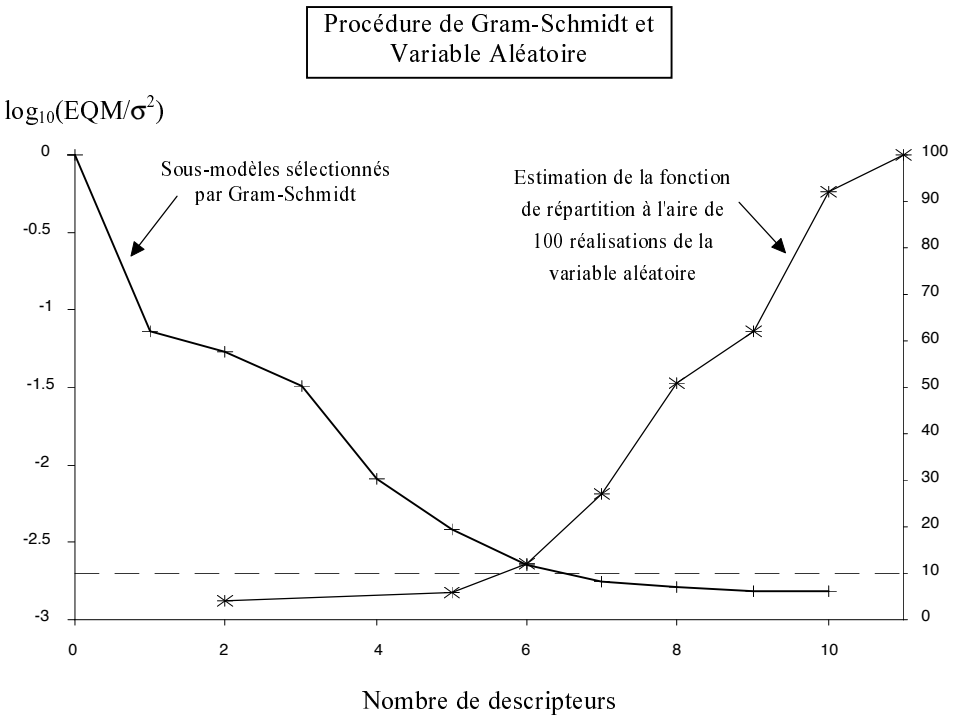


Figure 5.3 : Choix du sous-modèle avec une variable aléatoire

Comment lit-on la figure ?

- On se fixe dans un premier temps un niveau de probabilité, par exemple 10% (trait horizontal sur la figure 5.3).

- L'intersection de la courbe de répartition du classement de la variable aléatoire avec le niveau de probabilité fixé sélectionne le sous-modèle à 5 descripteurs.
- Ainsi, en sélectionnant le sous-modèle à 5 descripteurs, on peut dire que l'on a environ 10% de chance qu'un descripteur aléatoire explique mieux le problème posé qu'un des 5 descripteurs sélectionnés[3].

Notons que le sous-modèle sélectionné à l'aide de cette procédure (Gram-Schmidt et Variable Aléatoire) correspond au modèle "optimal" (modèle dont les paramètres sont tous non nuls, voir la construction de l'exemple au paragraphe 5.4.2.3). En revanche, le critère d'information d'Akaike (AIC(4)) sélectionne le sous-modèle à 6 descripteurs, sensiblement sur-dimensionné par rapport au modèle "optimal"[4].

L'intérêt de cette procédure est de montrer, plus concrètement que d'autres méthodes, la pertinence (ou l'absence de pertinence) de certains descripteurs par rapport à un descripteur aléatoire. Dans le paragraphe suivant, nous montrons comment les réalisations de la variable aléatoire peuvent être remplacées par le calcul de la distribution de probabilité de l'angle entre un vecteur aléatoire et le vecteur de sortie.

5.4.4 Calcul de la distribution de probabilité de l'angle entre le vecteur de sortie et un vecteur aléatoire

A chaque itération de la procédure de Gram-Schmidt, nous évaluons la proportion de vecteurs aléatoires qui font avec le vecteur de sortie un angle plus petit que celui que fait l'entrée sélectionnée avec le vecteur de sortie. Dans le paragraphe précédent, cette évaluation se faisait en engendrant plusieurs réalisations d'une variable aléatoire, puis en comptant celles dont l'angle avec le vecteur de sortie est plus faible. Nous allons montrer qu'il est possible de calculer exactement cette proportion, à partir de la répartition théorique du carré du cosinus de l'angle entre un vecteur aléatoire et un vecteur fixe.

L'ensemble des calculs est présenté dans l'Annexe B (Répartition de la variable aléatoire). La figure 5.4 représente la forme de la fonction de répartition du carré du cosinus entre un vecteur aléatoire et un vecteur fixe, de dimension N (N = nombre d'exemples, $N \geq 2$).

Cette fonction de répartition est notée :

$$fr_N\left(\cos^2(\theta)\right) \text{avec } N \geq 2$$

---

[3] Plus le niveau de probabilité est grand, plus le sous-modèle sélectionné est de grande taille, puisque la probabilité de garder un descripteur moins pertinent qu'une variable aléatoire est plus grande.

[4] Le principal défaut du critère d'information d'Akaike est d'être complètement faussé lorsque la valeur de l'EQM tend vers 0 (par exemple, lorsque le nombre de descripteurs et le nombre d'exemples sont du même ordre de grandeur).

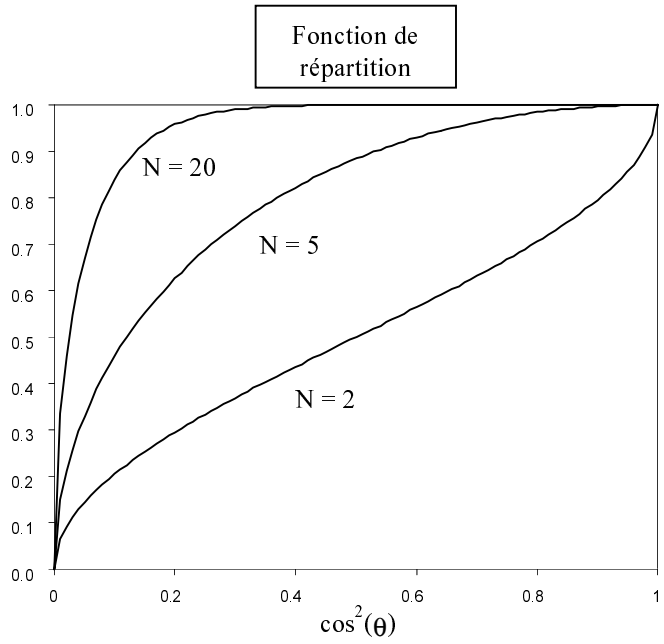


Figure 5.4 : Fonction de répartition pour N = 2, 5 et 20

Soit θ l'angle entre le descripteur sélectionné et le vecteur de sortie. Par définition de la fonction de répartition, la probabilité pour que le descripteur aléatoire explique mieux (angle plus petit) la sortie que le descripteur sélectionné, c'est-à-dire qu'il fasse avec la sortie un angle inférieur à θ, est donnée par l'expression suivante :

$$P_N\left(\cos^2(\theta)\right) = 1 - fr_N\left(\cos^2(\theta)\right) \text{ avec } N \geq 2$$

Pour illustrer la figure 5.4, nous considérons, par exemple, que le vecteur du descripteur sélectionné et le vecteur de sortie forment un angle d'environ 40 degrés.

Ainsi, nous avons :

$$\theta = 40° \text{ et } \cos^2(\theta) \approx 0{,}6$$

Avec un problème à 2 exemples (N = 2), nous lisons sur la Figure 5.4 que la fonction de répartition est égale à 0,55 ; soit une probabilité de 0,45. Ainsi, la probabilité qu'un descripteur aléatoire explique mieux la sortie que le descripteur considéré est de 45 %.

Pour un problème à 5 exemples (N = 5), cette probabilité tombe à 5 %. Elle devient quasiment nulle pour N = 20.

En pratique, ce calcul permet de s'affranchir des réalisations de variables aléatoires, et de l'application de la procédure de Gram-Schmidt à celles-ci : on effectue le classement des seuls P descripteurs par Gram-Schmidt ; pour le p-ième descripteur classé, on connaît $\cos^2\theta_p$, d'où l'on déduit la probabilité $P_{N\text{-}p}(\cos^2\theta_p)$ pour qu'un vecteur aléatoire fasse avec la sortie un angle inférieur à θ_p dans un espace de dimension N-p.

Rappelons que l'objectif est de déterminer la fonction de répartition du classement du descripteur aléatoire, c'est-à-dire la probabilité pour qu'un descripteur aléatoire soit classé dans un rang inférieur à p

Nous montrons dans l'annexe B que cette fonction de répartition est obtenue par la relation de récurrence suivante :

$$H_p = H_{p-1} + P_{N-p}\left(\cos^2 \theta_p\right) \cdot (1 - H_{p-1})$$

avec $H_0 = 0$

et H_p : probabilité pour qu'un descripteur aléatoire soit classé dans un rang inférieur à p, c'est-à-dire pour qu'un des p descripteurs sélectionnés soit moins significatif qu'un descripteur aléatoire.

La suite $\{H_p\}$ est croissante et bornée entre 0 et 1. Elle correspond à la probabilité d'avoir parmi les p descripteurs sélectionnés un descripteur ayant une contribution moindre que celle d'un descripteur aléatoire.

La figure 5.5 présente les répartitions de classement d'un descripteur aléatoire obtenues :

- soit à partir de 100 réalisations de la variable aléatoire,
- soit par la suite $\{H_p\}$ de la répartition de la variable aléatoire calculée comme indiqué dans l'annexe B.

On constate que les deux courbes sont proches (surtout pour les petits nombres de descripteurs). Ensuite, ces courbes s'éloignent car le nombre de réalisations de la variable aléatoire diminue, de sorte que l'estimation de la probabilité devient moins précise.

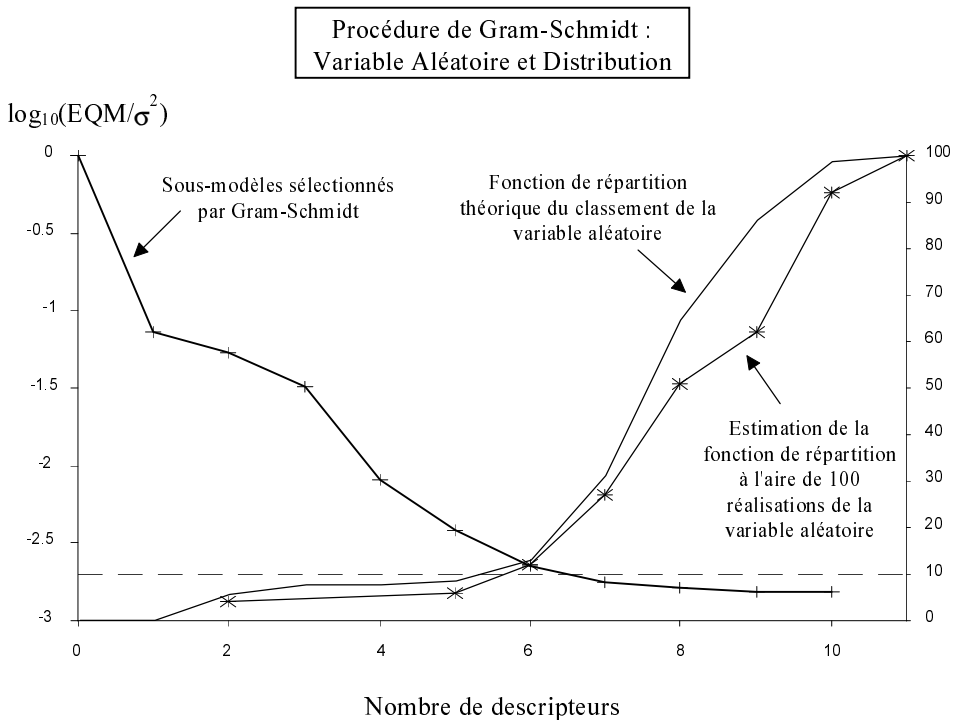


Figure 5.5 : Répartition estimée et théorique de la variable aléatoire

5.4.5 Mise en œuvre

La figure 5.6 présente l'organigramme reprenant les différentes étapes de la procédure originale de sélection de modèle. Avant de commencer la procédure, il faut fixer le niveau de probabilité pour qu'un descripteur aléatoire explique mieux la sortie qu'un descripteur sélectionné ; typiquement on prendra 5%.

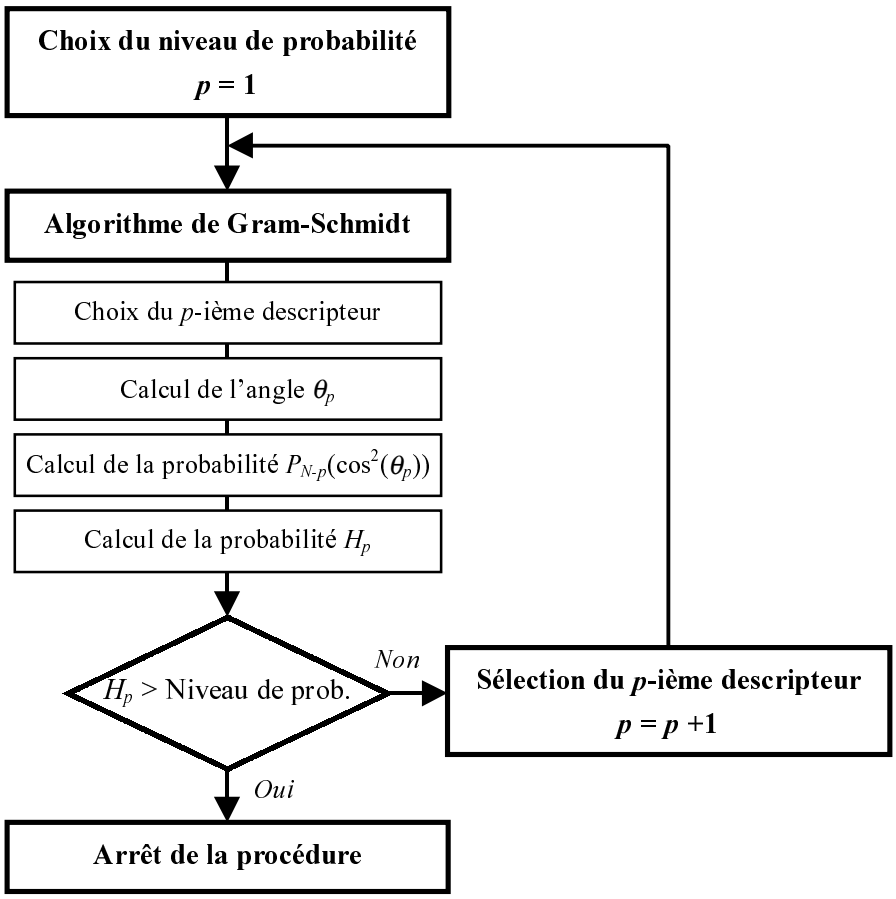


Figure 5.6 : Organigramme de la procédure de sélection de modèles

5.4.6 En résumé

La procédure de sélection de modèles proposée s'applique à la sélection de modèles linéaires par rapport aux paramètres ; elle s'appuie sur l'algorithme d'orthogonalisation de Gram-Schmidt, qui :

- ordonne les descripteurs suivant leur ordre d'importance,
- classe simultanément parmi ces descripteurs, plusieurs réalisations d'une variable aléatoire.

On a également vu que l'on pouvait se passer des réalisations de la variable aléatoire en calculant une fois pour toutes sa distribution. Ainsi, en se fixant un niveau de probabilité[5], cette procédure permet de sélectionner un sous-modèle parmi P sous-modèles.

Elle présente l'avantage de nécessiter un nombre de calculs réduit, et d'avoir une interprétation "intuitive".

---

[5] Lorsque l'on ne peut pas faire l'hypothèse d'un modèle linaire par rapport aux paramètres, on peut augmenter le niveau de probabilité afin de garder plus de descripteurs. Par exemple, on choisira un niveau de probabilité égal à 20% pour la sélection des descripteurs.

Elle ne peut pas s'appliquer directement à la sélection des entrées de réseaux de neurones multicouches, puisque leurs sorties ne sont pas linéaires par rapport aux poids de la première couche. Néanmoins, on peut sélectionner les descripteurs en utilisant un modèle linéaire par rapport aux paramètres, par exemple un modèle polynomial. Ensuite, on utilise les descripteurs sélectionnés comme entrées d'un réseau de neurones. Cette procédure a été mise en œuvre avec succès dans [Duprat 97]. L'annexe E (A New Decision Criterion for Feature Selection) présente un deuxième exemple d'utilisation de cette méthode concernant un dispositif embarqué de détection et de reconnaissance des défauts de rail débouchants.

Nous verrons dans le paragraphe suivant que cette procédure peut, en revanche, s'appliquer à la détermination automatique du nombre de neurones cachés dans un réseau de neurones statique à une couche cachée.

5.5 Détermination automatique de l'architecture d'un réseau de neurones

Nous avons décrit, dans le chapitre 3, les réseaux de neurones à une couche cachée qui permettent d'approcher toute fonction de régression, puis, dans le chapitre 4, nous avons présenté les algorithmes nécessaires à leur apprentissage. Nous abordons à présent le problème du dimensionnement d'un tel réseau, c'est-à-dire celui de la détermination du nombre de neurones cachés. Nous allons montrer que la procédure de sélection de modèles proposée précédemment peut être appliquée à ce problème.

Nous supposerons que les entrées du réseau ont été préalablement définies. Nous ne nous intéressons ici qu'au choix du nombre de neurones cachés.

5.5.1 Principe

Nous partons d'un réseau de neurones à une couche cachée et à sortie linéaire, pour lequel le nombre de descripteurs (nombre d'entrées) a été déterminé. L'idée est d'appliquer la méthode de sélection de modèles aux neurones cachés, dont les sorties constituent les entrées d'un "modèle" linéaire par rapport aux poids de la seconde couche de connexions. On effectue un premier apprentissage avec un nombre de neurones trop grand, puis, à l'aide de la méthode précédente, on élimine les neurones qui n'ont pas une contribution significative. On continue la procédure en poursuivant l'apprentissage avec les neurones restants, et en éliminant une nouvelle fois les neurones inutiles. On arrête ces itérations lorsque la procédure n'élimine plus aucun neurone.

5.5.2 Utilisation d'un "neurone aléatoire"

La figure 5.7 reprend la procédure de sélection de modèles en introduisant, non plus une entrée aléatoire, mais un "neurone aléatoire".

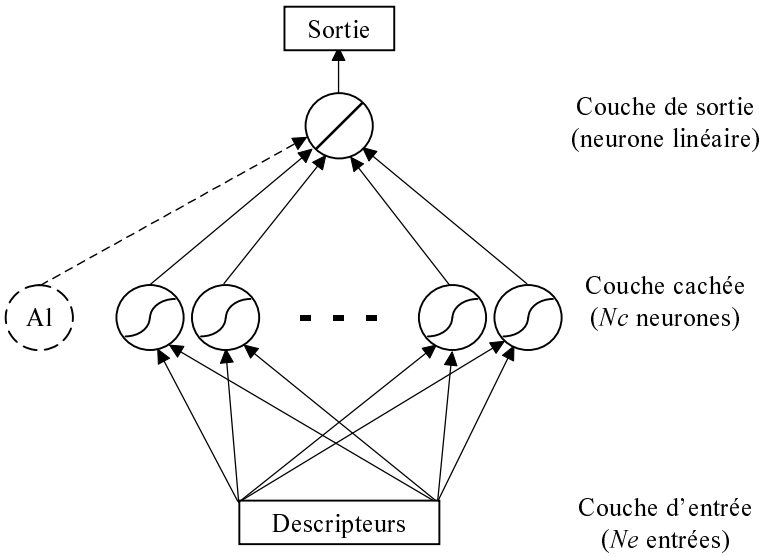


Figure 5.7 : Réseau de neurones à une couche cachée, à sortie linéaire, avec un neurone aléatoire

En fait, le neurone aléatoire n'existe pas dans la structure du réseau : après l'apprentissage, il suffit de considérer les sorties des neurones cachés comme les descripteurs d'un "modèle", constitué du neurone de sortie, qui est linéaire par rapport aux paramètres de la seconde couche de connexions. La procédure permet ensuite d'ordonner puis d'éliminer les neurones inutiles. On l'arrête lorsque que le "neurone aléatoire" se classe après les neurones cachés[6].

5.5.3 Mise en œuvre

Dans un premier temps, nous supposerons que les entrées du réseau ont été préalablement définies. Si ce n'est pas le cas, nous pouvons appliquer la procédure de sélection de modèles, comme indiqué ci-dessus.

Ensuite, il faut construire le modèle en choisissant un réseau de neurones de dimension appropriée.

De façon pratique, pour obtenir de bons résultats, il faut prendre quelques précautions. En effet, en cas de sur-apprentissage, le réseau obtenu utilise de manière significative tous les neurones dont il dispose, donc le neurone aléatoire est, logiquement, classé en dernière position. Dans ce cas, la procédure s'arrête immédiatement et ne réduit pas la dimension du réseau. Pour pallier cet inconvénient, il faut interrompre l'apprentissage (méthode dénommée *"early stopping"*, [Bishop 95]) en partitionnant l'ensemble des exemples ($\boldsymbol{E}_E$) en deux sous-ensembles :

- la première partie, notée $\boldsymbol{E}_A$ (environ 90% de $\boldsymbol{E}_E$), sert de base d'apprentissage,
- et le complément, noté $\boldsymbol{E}_S$ (environ 10% de $\boldsymbol{E}_E$), permet d'arrêter l'apprentissage.

---

[6] En fait, on arrête la procédure lorsque la probabilité qu'un "neurone aléatoire" soit classé avant les neurones du réseau ne dépasse pas le niveau de probabilité préalablement choisi.

L'apprentissage s'effectue donc sur les exemples de $\boldsymbol{E}_A$; à chaque itération on conserve la valeur de l'écart quadratique moyen sur l'ensemble d'arrêt $\boldsymbol{E}_S$ (noté EQMS). On applique la procédure de sélection de modèles aux sorties des neurones cachés calculées pour :

- les exemples de l'ensemble $\boldsymbol{E}_A$,
- et les coefficients du réseau correspondant à la valeur minimale de l'EQMS.

De cette manière, on supprime les neurones classés après le "neurone aléatoire". Après cette sélection des seuls neurones "utiles", on poursuit l'apprentissage du réseau de neurones obtenu. Mais avant de relancer l'apprentissage, il est nécessaire de modifier les coefficients qui relient la couche cachée au neurone de sortie. Pour cela, on utilise la méthode des moindres carrés ordinaires puisque la fonction d'activation du neurone de sortie est linéaire.

Cette mise en œuvre de la détermination de l'architecture d'un réseau de neurones s'inspire de la stratégie "destructive" de sélection de modèles (voir § 5.3.2). Il faut noter que l'on peut utiliser cette méthode suivant une stratégie "constructive" en partant d'un réseau de neurones comportant peu de neurones cachés. A chaque itération, on ajoute un neurone caché et, après la phase d'apprentissage on évalue si celui-ci apporte une contribution significative en le comparant au neurone "aléatoire".

5.5.4 Autres méthodes de détermination de l'architecture d'un réseau de neurones

Dans le domaine des réseaux de neurones, des méthodes ont déjà été proposées pour déterminer automatiquement l'architecture "optimale" pour un problème posé. Ce sont des techniques d'élagage : on choisit dans un premier temps un réseau de neurones sur-dimensionné, puis on réalise l'apprentissage et enfin, on supprime (on annule leur valeur) les paramètres (ou coefficients) qui ont le moins d'importance. Il faut donc mesurer l'importance relative d'un paramètre par rapport aux autres ; plusieurs voies sont possibles :

- la plus simple consiste à évaluer l'importance d'un paramètre comme l'amplitude de celui-ci, soit $|\theta_i|$. Cette approche n'est pas fondée sur des bases théoriques solides et donne de mauvais résultats [Bishop 95] ;
- on peut également calculer l'accroissement de la fonction de coût obtenu en supprimant le paramètre considéré. Les coefficients correspondant aux plus faibles augmentations sont éliminés. Cette approche a donné lieu à différentes méthodes d'élagage :
  - Ainsi, on peut annuler la valeur d'un paramètre du réseau, puis calculer l'accroissement de la fonction de coût sur l'ensemble d'apprentissage. On supprime finalement le paramètre qui donne l'accroissement le plus petit. Malheureusement, cette méthode nécessite un nombre important de calculs.
  - Pour pallier cet inconvénient, nous pouvons faire un développement de la fonction de coût au second ordre. Ainsi, l'accroissement du coût (J) obtenu avec une modification $\delta\theta$ des paramètres est donné par :

$$\delta J = \nabla J \, \delta\theta^T + \frac{1}{2} \delta\theta^T \, H \, \delta\theta + o\left(\delta\theta^3\right)$$

avec ∇J : gradient

et H : Hessien

Comme on suppose que l'apprentissage est terminé, le terme du premier ordre est négligeable (on a atteint un minimum local, donc le gradient est nul) ; ainsi on a :

$$\delta J = \frac{1}{2} \delta\theta^T \, H \, \delta\theta + o\left(\delta\theta^3\right)$$

La méthode dénommée *Optimal Brain Damage (OBD)* suppose que la matrice du Hessien est diagonale (les termes non diagonaux sont annulés) [Le Cun 90]. L'accroissement de la fonction de coût correspondant à l'élimination du paramètre *i* est alors :

$$\delta J_i = \frac{1}{2} H_{ii} \, \theta_i^2$$

[Hassibi 93] a apporté une amélioration à cette méthode en supprimant l'hypothèse sur le Hessien, (*Optimal Brain Surgeon : OBS*). L'accroissement est alors :

$$\delta J_i = \frac{1}{2} \frac{\theta_i^2}{H_{ii}^{-1}}$$

En fait, même si l'approximation est meilleure pour OBS que pour OBD, il n'existe aucune justification théorique de sa validité loin du minimum[7]. Au contraire, avec des modèles non linéaires par rapport aux paramètres, tels que les réseaux de neurones, la surface de coût n'est pas quadratique.

En d'autres termes, les méthodes OBD et OBS apportent toutes les justifications nécessaires au développement limité au voisinage du minimum, mais sont ensuite utilisées sans précaution très loin du minimum.

En résumé, ces méthodes offrent généralement des performances insuffisantes car leurs conditions de mise en œuvre ne respectent généralement pas les hypothèses sur lesquelles elles sont fondées.

5.5.5 En résumé

La méthode de sélection d'architecture de réseau de neurones que nous avons proposé donne de bons résultats ; en effet, nous avons constaté que le nombre de neurones choisi de cette manière était toujours satisfaisant. Elle s'adapte bien aux niveaux de bruit de la sortie observée et au nombre d'exemples à notre disposition.

---

[7] Avec OBD et OBS, on calcule l'accroissement de la fonction de coût sur les hyperplans d'équation : $\theta_i = 0$.

5.6 Cas de la classification

Nous avons présenté les méthodes de sélection de modèles dans le cadre de la modélisation de processus, qui considère le vecteur de sortie comme le vecteur du phénomène observé.

Dans le cas de la classification à C classes (ou le vecteur des sorties désirées ne peut prendre que C valeurs), nous ne pouvons plus apporter les justifications nécessaires aux méthodes de sélection de modèles. Nous devons nous contenter de les utiliser en ayant conscience de cette réserve. De façon pratique, on utilise un neurone de sortie linéaire pour la sélection des descripteurs et des neurones cachés. Puis, lorsque l'architecture du réseau est définie, on remplace ce neurone de sortie linéaire par un neurone possédant une fonction d'activation sigmoïdale.

Une première approche consiste à décomposer systématiquement un problème de classification à C classes en plusieurs sous-problèmes à 2 classes. Ensuite, la résolution (choix des descripteurs, choix de l'architecture du réseau de neurones et estimation des paramètres) de ces sous-problèmes se fait indépendamment. Le chapitre 2 (Méthodes statistiques de classification) présente cette approche.

Pour illustrer ce propos, nous pouvons considérer l'exemple de la reconnaissance de chiffres manuscrits. Comme les 10 classes (les chiffres de 0 à 9) ne sont pas ordonnées dans l'espace des descripteurs (dans cet espace, on ne passe pas successivement de la classe 0 à la classe 1, de 1 à 2, ...), les méthodes de sélection de modèles ne s'appliquent pas au traitement global de ce problème. Il faut le décomposer en sous-problèmes à deux classes, que l'on traite de manière indépendante. On supprime ainsi la relation d'ordre entre les différentes classes.

On trouve dans [Cibas 96] la présentation d'autres méthodes de sélection de modèles qui s'intègrent dans le cadre de la modélisation ou de la classification, et qui s'appliquent aux modèles linéaires ou non linéaires par rapport aux paramètres [voir également Leray 97].

5.7 Exemples d'application

Nous présentons trois exemples fictifs d'application de la méthode de sélection de modèles. Les deux premiers sont des problèmes de classification. Le premier porte sur la sélection des descripteurs, tandis que le second traite de la recherche d'une architecture adéquate pour un réseau de neurones. Nous présentons ensuite un exemple d'utilisation de procédure pour la modélisation.

5.7.1 1er exemple : Sélection des descripteurs

La réussite ou l'échec à un examen est un problème de classification à deux classes : "admis" ou "recalé". Prenons une population de 500 élèves ayant passé le baccalauréat l'année dernière et admettons que nous ayons la base d'apprentissage (imaginaire) suivante :

Élève	Note	Âge	Admis
Dupont	10.1	18.5	Oui
Durant	8.4	17.6	Non
...	...	...	...
500 exemples	2 descripteurs		2 classes

Tableau 5.1 : Base d'apprentissage

Le premier descripteur représente la moyenne pondérée des notes obtenues par l'élève à l'examen, le second est son âge. Nous nous trouvons ici dans une situation favorable à l'utilisation de méthodes de classification statistiques, car l'échantillon est de grande taille par rapport au nombre de descripteurs. Représentée dans l'espace des deux descripteurs (axe horizontal : note et axe vertical : âge), la population de l'échantillon apparaît ainsi :

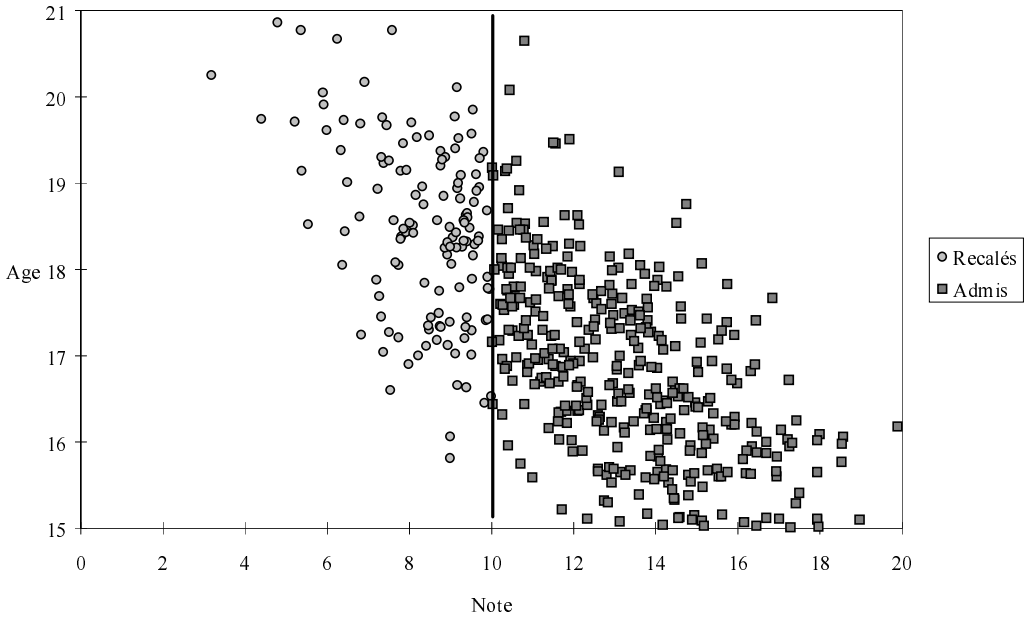


Figure 5.8 : Visualisation de l'échantillon

D'un coup d'œil, cette représentation confirme ce que tout le monde sait : ont leur bac tous les élèves qui ont au moins 10 de moyenne, et ce, quel que soit leur âge ! En effet, cette loi se voit sur le graphique puisque les deux classes sont parfaitement séparées par la droite verticale (note = 10) : les individus dont la note est supérieure ou égale à 10 sont admis et les autres sont recalés[8]. Les classes sont dites linéairement séparables.

[8] Avec ce problème à 2 dimensions (note et âge), nous sommes capables de représenter les individus dans le plan des descripteurs et constater qu'un seul descripteur (note) permet une séparation linéaire. C'est donc un descripteur pertinent pour caractériser l'appartenance des individus aux classes retenues. Face à un problème réel décrit par plus de 2 descripteurs, la représentation graphique devient impossible et il est beaucoup plus difficile d'identifier les descripteurs pertinents.

Nous allons confier ce problème à deux méthodes statistiques de résolution. Rappelons que ces méthodes s'efforcent de reproduire la classification de la base d'apprentissage ; elles sont d'autant meilleures qu'elles font moins d'erreurs de classement des individus[9]. La séparation des classes nous semblant ici très simple, on s'attend à ce que toutes les méthodes donnent de bons résultats. Nous allons constater que ce n'est pas le cas.

5.7.1.1 Analyse discriminante avec une règle d'affectation géométrique

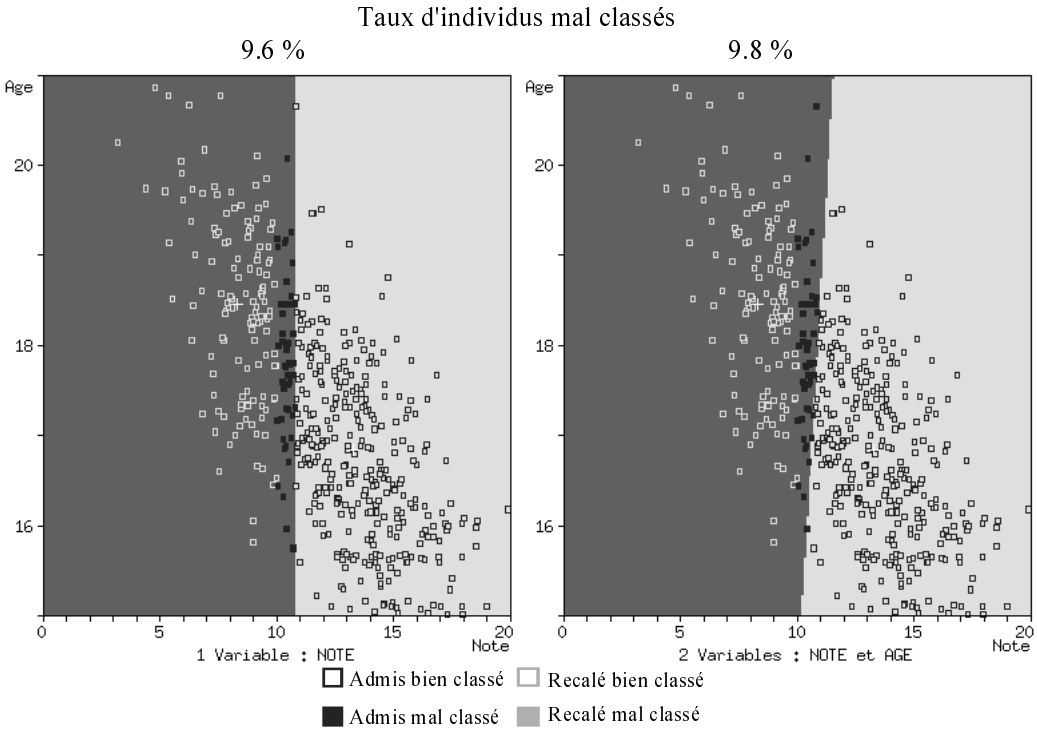


Figure 5.9 : Analyse discriminante avec une règle d'affectation géométrique

La zone dans laquelle le classifieur répond "admis" est caractérisée par un fond gris clair (gris foncé pour "recalé"). Les croix blanches sont les centres de gravités de deux classes.

Le graphique de gauche illustre le résultat obtenu en ne tenant compte que de la note comme descripteur. Le taux d'individus mal classés par la méthode est important (9,6%). La frontière de séparation est verticale et se situe aux environs de 10,8 donc assez loin du bon seuil. En s'appuyant sur les deux descripteurs note et âge (graphique de droite), l**es résultats se dégradent** : le taux d'individus mal classés passe à 9,8% et la frontière de séparation s'incline vers la droite.

---

[9] Dans la plupart des cas réels, il est cependant irréaliste d'espérer trouver une méthode qui fasse **aucune** erreur. Comme nous l'avons indiqué plus haut (voir chapitre 2), la règle de Bayes établit l'erreur minimale que l'on peut théoriquement espérer.

5.7.1.2 Réseaux de neurones

Ici, nous ne nous intéressons pas à la sélection de l'architecture du réseau de neurones mais seulement au choix des descripteurs : puisque les classes sont linéairement séparables, nous avons choisi un réseau de neurones constitué d'un seul neurone avec une fonction d'activation sigmoïde[10], de sorte que le problème de la sélection d'architecture ne se pose pas.

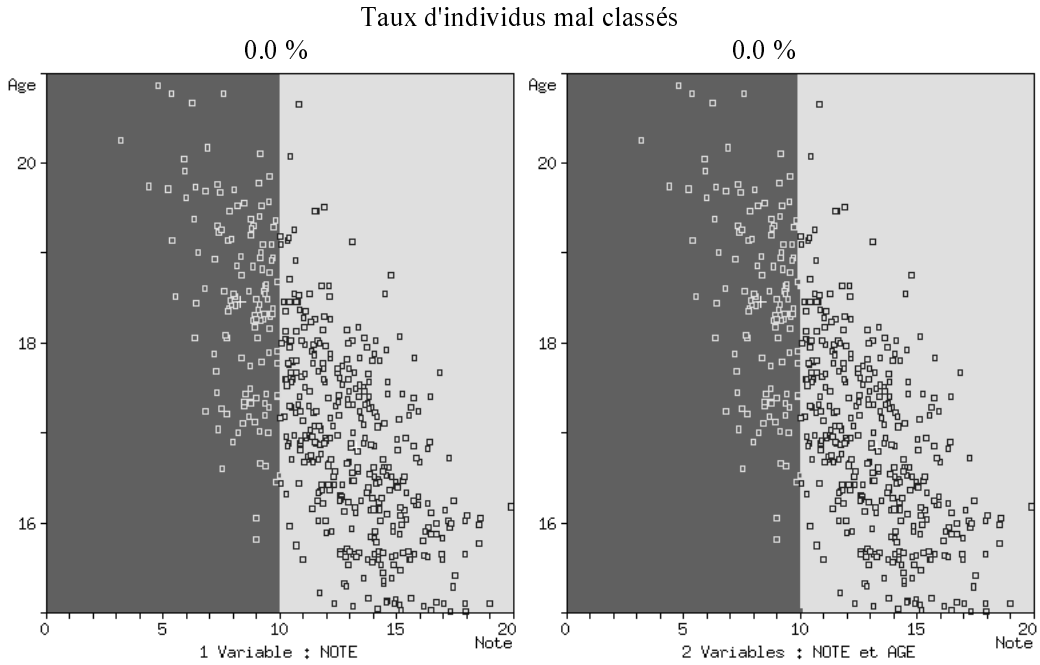


Figure 5.10 : Réseaux de neurones

La méthode ne fait aucune erreur de classement dans les deux cas (un descripteur ou deux). Le réseau a bien retrouvé la condition d'admission. Il faut nuancer ce résultat car, avec un problème moins bien construit (moins d'individus), cette méthode mène à un taux d'individus mal classés différent de 0% lorsqu'on tient compte des deux descripteurs (note et âge). On distingue, d'ailleurs, que la frontière de séparation n'est pas parfaitement verticale sur le graphique de droite (avec les 2 descripteurs).

Bien entendu, il faut souligner que la méthode de sélection de modèles proposée dans ce chapitre conduit à ne garder que la note comme descripteur pertinent (avec un seuil de probabilité de 20%).

5.7.1.3 Discussion

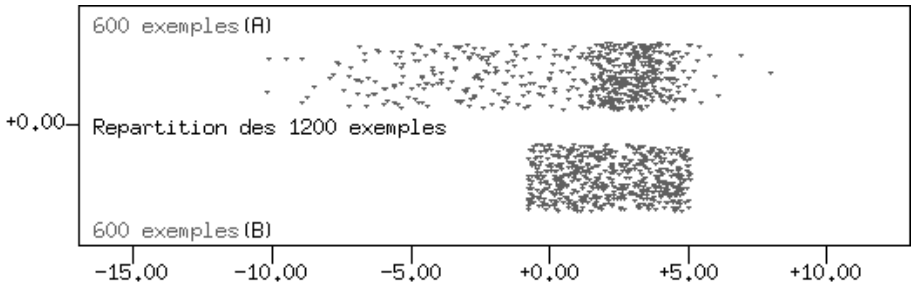
Sur cet exemple d'école, il apparaît que, dans tous les cas, le choix des descripteurs a une grande influence sur la qualité des résultats. Le "bruit" engendré par les descripteurs non pertinents dégrade gravement les résultats de classification obtenus par les méthodes statistiques. Ainsi, si une méthode donne de mauvais résultats, elle n'est pourtant pas

[10] Un tel réseau de neurones est suffisant pour résoudre un problème de classification présentant des classes linéairement séparables.

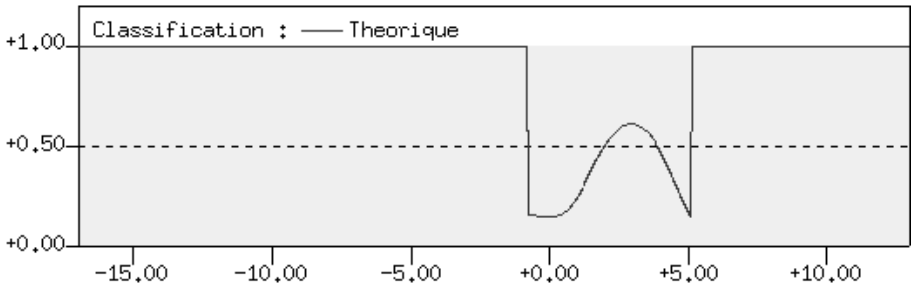
forcément à rejeter. Il se peut en effet que les descripteurs aient été mal choisis, et que, certains d'entre eux ne soient pas pertinents pour le problème posé.

5.7.2 2ème exemple : Architecture du réseau de neurones

Nous reprenons ici l'exemple donné au chapitre 2 (Méthodes statistiques de classification, § 2.5). Pour mémoire, la figure 5.11 présente cet exemple de classification à deux classes.



a/ Visualisation des 1200 individus



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

Figure 5.11 : Exemple de classification à une variable descriptive

Ici, l'objectif est de définir l'architecture du réseau de neurones (la sélection des descripteurs n'est pas abordée car ceux-ci sont tous pertinents).

5.7.2.1 Résultats

En partant d'un réseau de neurones comportant 20 neurones cachés, la méthode de sélection de modèles appliquée au choix des neurones cachés conduit à un réseau de neurones à 5 neurones cachés. Pour évaluer la qualité de cette réponse, nous avons essayé toutes les configurations possibles de réseaux de neurones (de 1 à 20 neurones cachés). Ici, nous représentons les meilleurs résultats obtenus en utilisant trois réseaux de neurones (4, 5 et 6 neurones cachés).

Le tableau 5.2 présente les résultats[11] :

[11] Rappel : la probabilité d'erreur de classification donné par la règle de Bayes est égal à 30,1%.

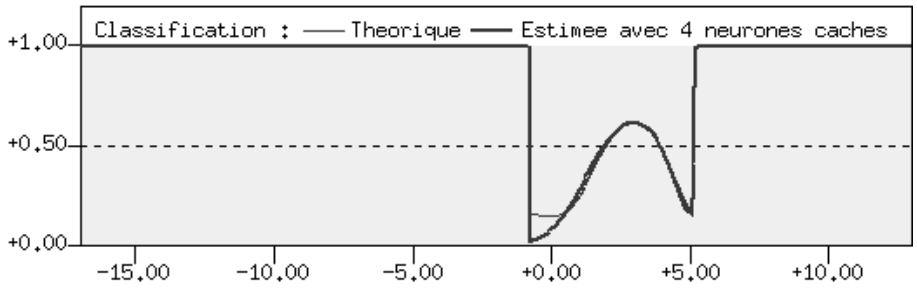
Nombre de neurones cachés	Taux d'exemples mal classés	Écart Quadratique Moyen
4	30,3%	$2{,}2\ 10^{-3}$
5	**30,3%**	**$1{,}2\ 10^{-3}$**
6	30,6%	$3{,}2\ 10^{-3}$

Tableau 5.2 : Résultats obtenus avec 3 réseaux de neurones

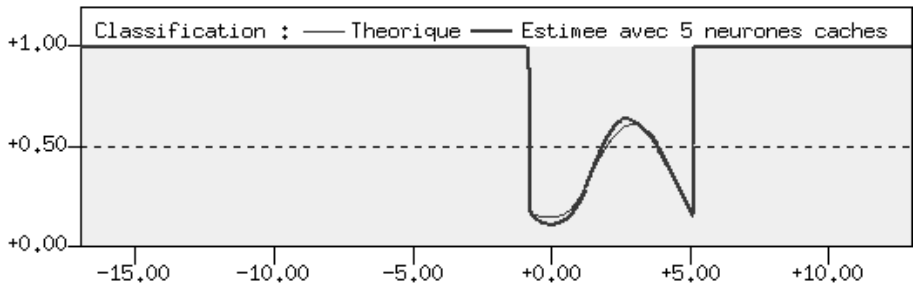
Dans le tableau nous présentons successivement :

- l'architecture du réseau de neurones considéré,
- le taux d'exemples mal classés,
- et l'écart quadratique moyen entre la probabilité *a posteriori* estimée par le réseau de neurones et la probabilité *a posteriori* théorique donnée par la règle de Bayes. Cet écart est estimé en générant plus d'un million de points (suivant les lois de distribution).

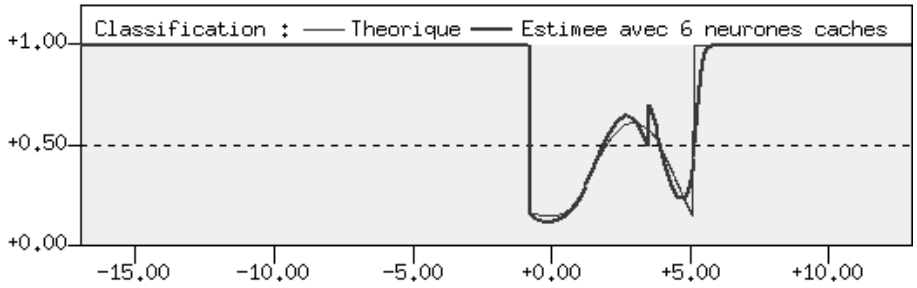
La figure 5.11 montre la probabilité *a posteriori* d'appartenance à la classe A estimée par les 3 réseaux de neurones d'architecture différente.



a/ Réseau de neurones à 4 neurones cachés



b/ Réseau de neurones à 5 neurones cachés



c/ Réseau de neurones à 6 neurones cachés

Figure 5.11 : Estimations obtenues avec les 3 réseaux de neurones

5.7.2.2 Discussion

La méthode de sélection de modèles conduit, de façon automatique, au meilleur réseau de neurones (à 5 neurones cachés). Elle présente donc un grand intérêt pratique pour résoudre toutes sortes de problèmes de classification.

5.7.3 3ème exemple : Problème "maître/élève"

Nous reprenons ici le problème de type "maître/élève" présenté au chapitre 4 (Apprentissage des réseaux de neurones).

Contrairement au chapitre 4 où le réseau "élève" possède la même architecture que le réseau "maître" (le problème est alors d'atteindre le minimum global de la fonction de coût), ici l'objectif est de retrouver l'architecture du réseau "maître" en partant d'un réseau "élève" sur-dimensionné. L'architecture à une couche cachée du réseau "maître" est présentée dans le tableau 5.3. Avec ce réseau "maître", on crée 2000 exemples d'apprentissage (voir chapitre 4).

Pour exécuter la procédure complète de détermination d'architecture de réseaux de neurones (sélection des descripteurs et des neurones cachés), on choisit l'architecture initiale du réseau "élève" suivante :

Les 10 descripteurs supplémentaires du réseau "élève" sont 10 réalisations différentes d'une variable aléatoire gaussienne (centrée et normée).

5.7.3.1 Résultats

La première phase de la procédure sélectionne correctement les 10 descripteurs du réseau "maître" (avec un niveau de probabilité égal à 5%).

Ensuite, la seconde phase s'arrête lorsqu'il ne reste plus que 5 neurones cachés (avec un niveau de probabilité égal à 5%). On retrouve correctement le bon réseau de neurones "maître".

Le tableau 5.3 présente l'architecture du réseau "élève" final :

	Nb descripteurs	Nb neurones cachés
Réseau de neurones "maître"	10	5
Réseau de neurones "élève" initial	20	10
Réseau de neurones "élève" final	10	5

Tableau 5.3 : Architectures du réseau "maître" et du réseau "élève" final

La procédure de détermination automatique de l'architecture d'un réseau de neurones conduit à un réseau "élève" identique au réseau "maître".

5.7.3.2 Discussion

Contrairement aux deux premiers exemples de classification, ce dernier exemple d'utilisation de la procédure de détermination de l'architecture d'un réseau de neurones s'insère dans le cadre de la modélisation de processus. Là encore, la méthode donne de bons résultats puisqu'elle a correctement retrouvé l'architecture du réseau de neurone "maître".

5.8 Conclusion

Dans ce chapitre, nous avons décrit diverses méthodes de sélection de modèles, et proposé une méthode originale de sélection de modèles. Cette méthode est intéressante car elle est économe en temps de calcul et met bien en évidence la pertinence des différents descripteurs. De plus, nous l'utilisons pour la sélection de l'architecture d'un réseau de neurones à une couche cachée.

Ainsi, face à un problème de modélisation ou de classification, la procédure permet :

- de définir les descripteurs,
- puis de déterminer le nombre de neurones de la couche cachée. À ce stade, le modèle est linéaire par rapport aux paramètres et les hypothèses sont donc vérifiées.

Ces deux traitements peuvent se réaliser presque automatiquement. Il suffit, pour cela, de se fixer les niveaux de probabilité.

En traitant différents exemples d'application, nous avons constaté que la démarche proposée (sélection des descripteurs pertinents et choix de l'architecture du réseau de neurones) permet d'améliorer les performances des classifieurs et, plus généralement des méthodes de modélisation.