TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

Détection et Estimation Supervisées[1]

P.COMON et G.BIENVENU

THOMSON - SINTRA
Parc de Sophia Antipolis, BP 138, 06561 Valbonne Cedex

Résumé

Les problèmes de détection et d'estimation supervisées reviennent à approcher une application d'un ensemble $\mathbb{E}$ dans un ensemble $\mathbb{F}$ à partir de sa valeur approximative en un nombre fini de points. Nous montrons que le choix du critère d'optimisation est crucial dans cet apprentissage. En particulier, la minimisation de l'erreur quadratique sur la base d'apprentissage est maladroite, bien qu'il soit montré ici que ses performances ultimes sont celles de la solution bayesienne.

Mots clés: Apprentissage supervisé, Classification, Bayesien, Maximum de vraisemblance, Neurone.

Abstract

Problems of supervised detection and estimation amount to the design of a mapping from a set $\mathbb{E}$ onto a set $\mathbb{F}$ based on its approximate value on a finite number of points. The choice of the optimization criterion is crucial. The use of quadratic error minimization over the learning set is awkward, although it is shown here that it would ultimately perform as well as the bayesian solution.

Key words: Supervised learning, Classification, Bayesian, Maximum Likelihood, Neuron.

1. Introduction

La détermination d'un récepteur optimal pour la détection ou l'estimation revient à construire une application φ de l'ensemble $\mathbb{E}$ des observations (que l'on suppose ici borné) dans un ensemble $\mathbb{F}$. Pour les problèmes de classification, l'ensemble $\mathbb{F}$ est discret, et son cardinal K correspond au nombre de classes possibles. Dans l'approche bayesienne, on s'intéresse plutôt à l'estimation des probabilités d'appartenance à chacune des classes, de sorte que l'ensemble $\mathbb{F}$ est $[0\ 1]^K$. Pour les problèmes d'estimation de paramètres déterministes ou aléatoires, les sorties prennent leurs valeurs dans un ensemble continu. Le problème étudié ici est l'identification ce cette application à partir d'exemples, en d'autres termes de façon supervisée. Ce mode d'apprentissage n'est pas le propre des réseaux de neurones, comme nous allons le montrer.

Nous présentons d'abord dans la section 2 un algorithme simple capable de construire le récepteur optimal au sens bayésien lorsque la base d'apprentissage est finie. La section 3 constitue le coeur de cet article. Nous y présentons des résultats asymptotiques démontrant l'inadéquation du critère d'erreur quadratique, souvent utilisé dans l'apprentissage des réseaux de neurones, et notamment du Perceptron Multicouche (MLP). Notre théorème I correspond à un résultat paru récemment dans [5], et obtenu indépendamment. Nos théorèmes II à IV le généralisent. La section 4 concerne le cas de paramètres continus, donc les problèmes d'estimation supervisée.

2. Apprentissage d'une densité de probabilité

2.1. Estimateurs nucléaires de densité

Les performances des estimateurs nucléaires sur des échantillons courts sont remarquables lorsque la densité cherchée est C^∞. Les propriétés fondamentales de ces estimateurs sont démontrées dans [3], tandis que [4] en fait un tour d'horizon synthétique.

Soit $\{x(n),\ 1\le n\le N,\ x(n)\in\mathbb{R}^M\}$ une base d'apprentissage de taille N, que l'on suppose être la réalisation d'une variable aléatoire unique, de densité $p_x(\mathbf{u})$. Alors on définit l'estimateur à noyau radial, $\hat{p}_x(\mathbf{u})$ comme suit.

$$(2\text{–}1)\qquad \hat{p}_X(N,\mathbf{u}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h(N)^M} K\left(\frac{\mathbf{u}-x(n)}{h(N)}\right),$$

où K est une fonction de $\mathbb{R}^M$ dans $\mathbb{R}$ ne dépendant que de la norme de son argument, et h(N) une suite réelle positive tendant vers zéro quand $N\to\infty$. Cacoullos [3] énonce un certain nombre de théorèmes exhibant les conditions que doit verifier h(N) pour que l'estimateur soit performant. En particulier:

Théorème de convergence

Si $K(\mathbf{u})$ et $p_x(\mathbf{u})$ vérifient les conditions suivantes

$$(2\text{–}2)\qquad \int_{\mathbb{R}^M} K(\mathbf{u})\, d\mathbf{u} = 1,$$

$$(2\text{–}3)\qquad \int_{\mathbb{R}^M} K(\mathbf{u})\, u_i\, u_j\, d\mathbf{u} = v_{ij} \text{ borné}, \ \forall\ i,j,\ 1 \le i,j \le M,$$

(2–4) $p_X(x)$ est de classe C^2, et de dérivée bornée,

et si h(N) vérifie

$$(2\text{–}5)\qquad \lim_{N\to\infty} h(N) = 0 \quad \text{et} \quad \lim_{N\to\infty} N\, h(N)^M = \infty,$$

alors l'estimateur (2–1) converge en moyenne quadratique; autrement dit, l'erreur

$$(2\text{–}6)\qquad e(N) = E\int_{\mathbb{R}^M} [\hat{p}_X(N,\mathbf{u})] - p_X(\mathbf{u})]^2\, d\mathbf{u}$$

tend vers zéro quand $N\to\infty$.

Cacoullos énonce même un résultat plus fort, à savoir que l'erreur e(N) admet une borne inférieure, et que cette borne est atteinte pour certaines fonctions h(N):

$$(2\text{–}7)\qquad e(N) = O(N^{-\frac{4}{M+4}}) \text{ pour } h(N) = O(N^{-\frac{1}{M+4}}).$$

Malheureusement, il est en général impossible de calculer le coefficient de proportionnalité liant $h(N)^{M+4}$ et N^{-1}. Nous allons proposer dans les deux paragraphes qui vont suivre une heuristique palliant cette difficulté.

2.2. Agglomération

Lorsqu'un nombre croissant d'échantillons x(n) sont présentés, il devient de toutes façons difficile de calculer la densité par la fomule

[1]: Ce travail a été en partie financé par la D.R.E.T., Paris.

(2–1), les contributions des $\mathbf{x}(n)$ étant trop nombreuses. L'agglomération automatique permet de surmonter cet obstacle. Plusieurs techniques peuvent être utilisées à cet effet. Nous suggérons de représenter l'ensemble de la base d'apprentissage A_0 par un nombre fixé P de "centroïdes", $\mathbf{w}(p)$, auquels on attribue des effectifs, a(p) [6]. Ainsi, l'ensemble A_0 est constitué de P groupes disjoints, E(p). Pour ce faire, chacun des vecteurs $\mathbf{x}(n)$ est affecté au groupe dont le centroïde est le plus proche [6].

2.3. Reconstruction de la densité

L'estimateur (2–1) peut se réécrire de la façon suivante:

$$(2\text{–}8)\qquad \hat{p}_X(N,\mathbf{u}) = \frac{1}{\sum a(p)} \sum_{p=1}^{P} \sum_{\mathbf{x}(n)\in E(p)} \frac{1}{h(a(p))^M} K\left(\frac{\mathbf{u}-\mathbf{x}(n)}{h(a(p))}\right),$$

où le facteur h a été autorisé à varier d'un groupe à l'autre. Afin de déterminer une heuristique permettant de calculer le coefficient h(a(p)), nous faisons plusieurs hypothèses simplificatrices. La première consiste à assimiler chaque échantillon à son centroïde. Ceci nous donne d'après (2–8)

$$(2\text{–}9)\qquad \hat{p}(\mathbf{u}) = \frac{1}{\sum a(p)} \sum_{p=1}^{P} \frac{a(p)}{h(a(p))^M} K\left(\frac{\mathbf{u}-\mathbf{w}(p)}{h(a(p))}\right).$$

En outre, nous admettons que la densité à reconstruire est une mixture de P gaussiennes isotropes, de moyenne $\mathbf{w}(p)$ et de variance $\sigma^2(p)$ I:

$$(2\text{–}10)\quad p(\mathbf{v}/\mathbf{v}\in E(p)) = (2\pi)^{-M/2}\, \sigma(p)^{-M} \exp(-\|\mathbf{v}-\mathbf{w}(p)\|^2 / 2\sigma(p)^2),$$

et nous choisissons pour K le noyau gaussien standardisé:

$$(2\text{–}11)\qquad K(\mathbf{v}) = (2\pi)^{-M/2} \exp(-\|\mathbf{v}\|^2 / 2).$$

Ce noyau satsifait les conditions que nous avons énoncées dans la section 2.1. Alors, en reprenant le détail des calculs de Cacoullos [3], on peut montrer que la valeur optimale de h vérifie:

$$h(a(p))_{opt}^{M+4} = M \int_{\mathbb{R}^M} K^2(\mathbf{z})\, d\mathbf{z} \Big/ a(p) \int_{\mathbb{R}^M} \Delta p(\mathbf{x}/\mathbf{x}\in E(p))^2\, d\mathbf{x},$$

où Δ désigne l'opérateur laplacien. En utilisant (2–10) et (2–11), cette relation conduit après calculs à:

$$(2\text{–}12)\qquad h(a(p))_{opt}^{M+4} = \frac{M}{a(p)} \frac{4}{3M\,\sigma(p)^{-M-4}},$$

soit

$$h(a(p))_{opt} = \sigma(p) \left(\frac{4}{3}\right)^{\frac{1}{M+4}} a(p)^{-\frac{1}{M+4}}.$$

Quant au terme $\sigma(p)$, il peut être remplacé par son estimation du maximum de vraisemblance:

$$(2\text{–}13)\qquad \hat{\sigma}(p)^2 = \frac{1}{M\,a(p)} \sum_{\mathbf{x}\in E(p)} \|\mathbf{x}-\mathbf{w}(p)\|^2.$$

En conclusion, les relations (2–12) et (2–13) nous fournissent une valeur de h plausible pour chaque groupe E(p), et les relations (2–8) ou (2–9) nous donnent alors une estimation de la densité. Si les moyens de calcul le permettent, (2–8) sera évidemment plus précise que (2–9).

3. Apprentissage d'un classifieur

3.1. Codage des sorties

Les performances de l'approche proposée peuvent être comparées à celles du Perceptron Multicouche (MLP, pour Multi Layer Perceptron) nanti de l'algorithme de Rétropropagation du Gradient très en vogue dans la communauté des réseaux neuronaux. L'architecture imposée par le MLP revient à chercher l'application φ dans une famille Φ fixée de fonctions paramétrées par un tableau de paramètres, $\mathbb{W}$. Notons $\varphi(\mathbb{W},\bullet)$ l'élément générique de Φ. Dans le MLP, cette fonction a une forme particulière, que nous allons passer sous silence car elle nous importe peu pour le propos qui va nous occuper. Retenons simplement que la famille Φ est dense dans l'espace des fonctions continues de $\mathbb{E}$ dans $\mathbb{F}$, pour la plupart des fonctions de transition neuronales, pourvu que $\mathbb{E}$ et $\mathbb{F}$ soient des ensembles fermés bornés [1] [2]. Nous supposerons dorénavant que c'est le cas, et que $\mathbb{E} \subset \mathbb{R}^M$, et $\mathbb{F} \subset \mathbb{R}^K$.

L'apprentissage d'un classifieur dans la famille Φ consiste à déterminer le vecteur $\mathbb{W}$, définissant complètement l'application $\varphi(\mathbb{W},\bullet)$ de $\mathbb{E}$ dans $\mathbb{F}$ définie par

$$(3\text{–}1)\qquad \mathbf{x}\in\mathbb{E} \rightarrow \mathbf{y} = \varphi(\mathbb{W},\mathbf{x}) \in \mathbb{F},$$

à partir des N exemples disponibles $\{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, 1\le n\le N\}$. La solution $\varphi(\mathbb{W},\bullet)$ obtenue dépend a priori du codage des sorties, de la fonction φ, et du critère d'optimisation choisis.

Parlons d'abord du codage des classes. Il existe une infinité de façons d'établir une correspondance entre K classes et des valeurs numériques. On pourrait par exemple choisir $\mathbb{F} = \{1,2,\ldots,K\}$. Ce choix est peu pratiqué. Une autre solution, la plus répandue, consiste à prendre pour ensemble $\mathbb{F}$ le sous-ensemble de $\{0,1\}^K$ des vecteurs de norme 1: $\mathbb{F} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$.

Dans ce cas, la fonction φ n'est autre que la fonction indicatrice des classes:

$$(3\text{–}2)\qquad \mathbf{x}^{(n)} \in \text{Classe } \omega_k \Leftrightarrow y_i^{(n)} = \delta_{ik},\ 1\le i\le K.$$

Par exemple la sortie associée à la classe 2 s'écrira $\mathbf{e}_2$=(0,1,0,...0).

3.2. Apprentissage quadratique

On pourrait croire que le codage ne change rien à l'affaire, et ce serait une erreur car cela dépend beaucoup du critère d'optimisation choisi. L'apprentissage du MLP est habituellement réalisé en minimisant l'erreur quadratique $\mathbf{y} - \varphi(\mathbb{W},\mathbf{x})$ pour l'ensemble des couples $(\mathbf{x},\mathbf{y})$ de la base d'apprentissage, soit:

$$(3\text{–}3)\qquad \mathbb{W} = \operatorname*{Arg\,Min}_{\mathbb{W}} \sum_{n=1}^{N} \|\mathbf{y}^{(n)} - \varphi(\mathbb{W},\mathbf{x}^{(n)})\|^2,$$

avec un codage donné, qui peut être par exemple (3–2).

Evidemment, à cause des nécessaires non linéarités de la fonction φ par rapport aux éléments de $\mathbb{W}$, cette minimisation peut soulever de grosses difficultés car il existe en général des minima locaux. Ici nous supposerons que le minimum absolu peut être toujours obtenu exactement, de sorte que nous évaluerons des performances "ultimes".

3.3. Apprentissage bayesien

Dans l'approche bayesienne, $\mathbb{F} = [0,1]^K$, et le critère d'optimisation est le risque probabiliste:

$$(3\text{–}4)\qquad R = \sum_{i,j=1}^{K} \kappa(i,j)\, P_j \int_{\mathbf{u}\in\omega_i} p(\mathbf{u}/\omega_j)\, d\mathbf{u},$$

où P_j désigne la probabilité a priori qu'une observation soit issue de la classe ω_j, $\kappa(i,j)$ le coût associé à la classification d'un élément de la classe ω_j dans la classe ω_i, et $p(\mathbf{u}/\omega_j)$ la probabilité conditionnelle qu'une observation issue de la classe ω_j prenne la valeur $\mathbf{u}$. Nous avons par exemple, avec ces notations, la relation donnant la densité de probabilité de l'observation:

$$(3\text{–}5)\qquad p(\mathbf{u}) = \sum_{k=1}^{K} P_k\, p(\mathbf{u}/\omega_k).$$

Soit $\mathbf{x}$ une observation présentée. Alors minimiser le risque R revient à affecter $\mathbf{x}$ à la classe ω_i pour laquelle la quantité

$$(3\text{–}6)\qquad B_i(\mathbf{x}) = \sum_{j=1}^{K} \kappa(i,j)\, P_j\, p(\mathbf{x}/\omega_j)$$

est la plus petite. Ceci découle directement de (3–5). Un cas particulier d'importance est celui des coûts uniformes: $\kappa(i,j) = 1-\delta_{ij}$. Dans ce cas la relation (3–5) permet d'écrire (3–6) sous une autre forme:

$$B_i(\mathbf{x}) = p(\mathbf{x}) - P_i\, p(\mathbf{x}/\omega_i).$$

Autrement dit, puisque $p(\mathbf{x})$ ne dépend pas de ω_i, on peut de manière équivalente affecter $\mathbf{x}$ à la classe pour laquelle la quantité

$$(3\text{–}7)\qquad b_i(\mathbf{x}) = P_i\, p(\mathbf{x}/\omega_i)$$

est maximale. En d'autres termes la *minimisation* de $B_i(\mathbf{x})$ ou la *maximisation* de $b_i(\mathbf{x})$ par rapport à i sont strictement équivalentes

lorsque les coûts sont uniformes. La matrice des coûts, de taille K^2, ne possède d'ailleurs que K^2–K–1 degrés de liberté, ce qui nous autorise en particulier à ajouter une constante aux fonctions discriminantes ou à les changer de signe [7]. Notons au passage que $b_i(\mathbf{x})$ n'est autre que la probabilité conjointe $p(\omega_i,\mathbf{x})$, probabilité d'observer un vecteur d'entrée $\mathbf{x}$ issu de la classe ω_i. En pratique, les densités conditionnelles peuvent être estimées à partir des estimateurs proposés dans les sections 2.1 et 2.3.

3.4. Performances ultimes pour des bases de taille infinie

Nous énonçons dans cette section des résultats asymptotiques tendant à montrer que même avec une infinité d'exemples, le critère quadratique ne pourra guère faire mieux que le critère bayesien.

Théorème I.

Soit $A_0=\{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, 1\le n\le N\}$ une base d'apprentissage, contenant N_k exemples dans chacune des classes ω_k, $1\le k\le K$. Soit l'application $\varphi(\mathbb{W},\bullet)$ obtenue en minimisant l'erreur

$$\varepsilon(N) = \frac{1}{N}\sum_{n=1}^{N} \| \mathbf{y}^{(n)} - \varphi(\mathbb{W}, \mathbf{x}^{(n)}) \|^2 \qquad (3\text{–}8)$$

par rapport à $\mathbb{W}$. Alors cette application tend lorsque $\{N_k \to \infty, 1\le k\le K\}$ vers la meilleure approximation quadratique du récepteur optimal bayesien équipénalisé (3–7) dans l'ensemble Φ, à condition de choisir le codage (3–2).

Théorème II.

Soit $\{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, 1\le n\le N\}$ une base d'apprentissage de taille finie N, contenant N_k exemples dans chacune des classes ω_k, $1\le k\le K$. Alors la solution $\varphi(\mathbb{W},\bullet)$ obtenue en minimisant l'erreur euclidienne

$$\varepsilon(N) = \frac{1}{N}\sum_{n=1}^{N} \| \mathbf{y}^{(n)} - \varphi(\mathbb{W}, \mathbf{x}^{(n)}) \|^2$$

tend vers la meilleure approximation dans Φ de la solution bayesienne générale (3–6) lorsque $\{N_k \to \infty, 1\le k\le K\}$, à condition de choisir pour codage

$$y_i^{(n)} = \kappa(i,j) \text{ pour } \mathbf{x}^{(n)} \in \omega_j. \qquad (3\text{–}9)$$

Ces deux théorèmes sont démontrés en annexe.

3.5. Performances ultimes pour des bases de taille finie

Lorsque N est fini, une méthode couramment pratiquée pour étendre artificiellement la taille de la base $A_0(N)=\{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, 1\le n\le N\}$ consiste à la dupliquer en la bruitant à l'aide de NQ réalisations indépendantes d'un bruit centré de densité $p_z(\mathbf{u})$. Nous obtenons ainsi Q bases d'apprentissage de taille N:

$$A_q(N) = \{\mathbf{x}^{(n)}+\mathbf{z}^{(n,q)}, \mathbf{y}^{(n)}, 1\le n\le N\},\ 1\le q\le Q. \qquad (3\text{–}10)$$

Posons alors

$$A(N,Q) = \bigcup_{q=1}^{Q} A_q(N). \qquad (3\text{–}11)$$

Théorème III

Soit $A_0(N)=\{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}, 1\le n\le N\}$ une base d'apprentissage de taille finie contenant N_k exemples ($N_k>0$) dans chacune des classes, et A(N,Q) sa version étendue conformément à (3–11). On suppose en outre que le codage adopté est celui de (3–2). Alors lorsque Q tend vers l'infini, la solution $\varphi(\mathbb{W},\bullet)$ obtenue en minimisant l'erreur euclidienne

$$\varepsilon(N,Q) = \frac{1}{N}\frac{1}{Q}\sum_{n=1}^{N}\sum_{q=1}^{Q} \| \mathbf{y}^{(n)} - \varphi(\mathbb{W}, \mathbf{x}^{(n)}+\mathbf{z}^{(n,q)}) \|^2 \qquad (3\text{–}12)$$

tend vers la meilleure approximation dans Φ de la solution bayesienne équipénalisée (3–7) construite à partir de l'estimateur nucléaire (2–1) de noyau $p_z(\mathbf{u})$.

Théorème IV

Dans les mêmes conditions que le théorème ci-dessus, et lorsque les sorties $\mathbf{y}^{(n)}$ sont définies par le codage (3–9), la solution obtenue en minimisant l'erreur $\varepsilon(N,Q)$ tend vers la solution bayesienne générale (3–6) où les densités sont remplacées par leurs estimations nucléaires de noyau $p_z(\mathbf{u})$.

Les démonstrations, données en annexe, exhibent des conditions portant sur le bruit $\mathbf{z}$ permettant à la solution d'être consistante. En effet, on y constate que les densités conditionnelles $p(\mathbf{v}/\omega_k)$ tendent vers leurs estimations nucléaires (2–1) de noyau

$$p_z(\mathbf{u}) = K(\mathbf{u}/h)/h^M. \qquad (3\text{–}13)$$

Par conséquent, grâce au théorème de Cacoullos, on constate qu'il convient de choisir un bruit additif $\mathbf{z}$ tel que sa densité soit définie par (3–13) où h vérifie (2–7).

Corollaire

Lorsqu'on créera les bases $A_q(N)$, on pourra notamment choisir un bruit gaussien isotrope d'écart-type $\sigma = \alpha\, N^{-1/M+4}$, où α est une constante fixée.

4. Apprentissage d'un estimateur

Le critère standard utilisé pour l'estimation de paramètres déterministes est la Maximisation de la Vraisemblance (MV); ce critère est en général difficile à mettre en oeuvre lorsque la densité de probabilité conditionnelle est multimodale, et de surcoît inconnue. Les mêmes difficultés existent dans les problèmes d'estimation de variables aléatoires, au sens du Maximum A Posteriori (MAP) par exemple. L'approche proposée consiste à identifier le récepteur optimal, $\varphi(\mathbf{x})$, à partir des exemples de la base d'apprentissage, $A_0=\{\mathbf{x}^{(n)},\mathbf{y}^{(n)},1\le n\le N\}$.

Plutôt que de suivre la même démarche qu'au chapître 3 en supposant que $\mathbf{y}^{(n)}\approx\varphi(\mathbf{x}^{(n)})$, nous supposerons que l'ensemble A_0 définit les coordonnées d'une ligne de crête d'une densité conditionnelle $p_x(\mathbf{x}/\mathbf{y})$, si $\mathbf{y}$ est déterministe, ou d'une densité conjointe $p(\mathbf{x},\mathbf{y})$ si $\mathbf{y}$ est aléatoire, ce que nous conviendrons de noter de la même façon dans les deux cas:

$$\mathbf{y}^{(n)} = \text{Arg}\ \underset{\mathbf{y}\in\mathbf{R}^K}{\text{Max}}\ p(\mathbf{x}^{(n)}, \mathbf{y}). \qquad (4\text{–}1)$$

Avec cette approche, nous pouvons mettre à profit les résultats du chapître 2. A partir des exemple données dans A_0, il est possible d'estimer la fonction $p(\mathbf{x},\mathbf{y})$ en utilisant l'estimateur (2–8). Lorsqu'une observation $\mathbf{x}$ sera présentée, on pourra obtenir l'estimation $\hat{\mathbf{y}}=\varphi(\mathbf{x})$ en cherchant le maximum absolu par rapport à $\mathbf{y}$ de la coupe $\hat{p}(\mathbf{x},\mathbf{y})$, à $\mathbf{x}$ fixé [8]. Comme le codage de cette densité a été fait à l'aide d'une agglomération préalable, nous disposons d'excellentes valeurs initiales que constituent les P centroïdes $\mathbf{w}(p)$, si nous voulons exécuter un algorithme d'ascente. Le maximum absolu sera obtenu en sélectionnant le plus grand des P maxima locaux ainsi obtenus.

Un des principaux avantages de cette procédure est la robustesse aux données lacunaires. En effet, si une composantes de $\mathbf{x}$, disons x_M par exemple, est manquante, il suffit de de faire une coupe de $\hat{p}(\mathbf{x},\mathbf{y})$ selon le vecteur $(x_1,\ldots,x_{M-1})$, et de rechercher son maximum par rapport aux variables $(x_M, y_1,\ldots,y_K)$ [8]. La procédure fournit donc en plus de la solution $\hat{\mathbf{y}}=\varphi(\mathbf{x})$ une estimation du paramètre d'entrée manquant, $\hat{x}_M$.

La solution bayesienne correspondant au coût quadratique peut aussi être calculée aisément à partir de la densité conjointe $\hat{p}(\mathbf{x},\mathbf{y})$ [8]. Pour ce faire il suffit d'intégrer la quantité $\mathbf{y}\,\hat{p}(\mathbf{x},\mathbf{y})$ sur l'espace du paramètre $\mathbf{y}$, pour l'observation $\mathbf{x}$ considérée.

Références

[1] K.HORNIK, "Multilayer Feedforward Networks are Universal Approximators", *Neural Networks*, vol.2, n°5, 1989, 359-365.

[2] G.CYBENKO, "Approximation by Superpositions of Sigmoidal Functions", *Mathematics of Control, Signals and Systems*, 2, n°4, 1989.

[3] T.CACOULLOS, "Estimation of a Multivariate Density", *Annals of Inst. of Stat. Math.*, 18, 1966, 178-189.

[4] D.J.HAND, *Kernel Discriminant Analysis*, 1982, RSP press.
[5] D.W.RUCK, S.K.ROGERS et al, "The Multilayer Perceptron as an Approximation to Bayes Optimal Discriminant Function", *IEEE Trans. Neural Networks*, vol.1, dec 1990, 296-298.
[6] P.COMON, "Classification Bayesienne Distribuée", *Revue Technique Thomson*, vol.22, n°4, dec 1990, 543-562.
[7] F.VALLET, "Approche Globale du Problème de Discrimination", *Revue Technique Thomson*, vol.22, n°4, dec 1990, 519-542.
[8] P.COMON, "Supervised Detection and Estimation", *ESPRIT-BRA Workshop on Neural Networks and Artificial Vision*, jan 29-30, 1991, Chamrousse.

Annexes

♦Lemme

Les démonstrations des théorèmes I et II débutent de la même façon. La définition de l'erreur donnée en (3–8) peut s'écrire

$$\varepsilon(N) = \sum_{k=1}^{K} \frac{N_k}{N} \frac{1}{N_k} \sum_{\mathbf{x}^{(n)} \in \omega_k} \| \mathbf{y}^{(n)} - \varphi(\mathbb{W}, \mathbf{x}^{(n)}) \|^2.$$

Faisons tendre les N_k vers l'infini, nous obtenons:

$$\varepsilon(\infty) = \sum_{k=1}^{K} P_k \int p(\mathbf{u}/\omega_k) \| \mathbf{y}^{(n)} - \varphi(\mathbb{W}, \mathbf{u}) \|^2 \, d\mathbf{u}.$$

Puis en utilisant la relation

$$p(\mathbf{u}) = \sum_{k=1}^{K} P_k \, p(\mathbf{u}/\omega_k),$$

nous obtenons finalement:

$$\text{(A–1)} \qquad \varepsilon(\infty) = \int p(\mathbf{u}) \, \| \varphi(\mathbb{W}, \mathbf{u}) \|^2 \, d\mathbf{u} - 2 \sum_{k=1}^{K} P_k \int p(\mathbf{u}/\omega_k) \sum_{j=1}^{K} y_j^{(n)} \varphi_j(\mathbb{W},\mathbf{u}) \, d\mathbf{u} + \varepsilon_0,$$

où ε_0 est indépendant des φ_j. ◊

♦Démonstration du théorème I

Ici, $y_j^{(n)} = \delta_{jk}$ si $\mathbf{x}^{(n)} \in \omega_k$. Dans (A–1), la somme sur j se réduit donc à un seul terme, $\varphi_k(\mathbb{W},\mathbf{u})$. Posons $g_k(\mathbf{u}) = b_k(\mathbf{u})/p(\mathbf{u})$, où $b_k(\mathbf{u})$ a été défini en (3–7); remarquons que $g_k(\mathbf{u})$ n'est autre que la probabilité conditionnelle de la classe k, $p(\omega_k/\mathbf{u})$. L'expression (A–1) devient

$$\text{(A–2)} \qquad \varepsilon(\infty) = \int p(\mathbf{u}) \, \| \varphi(\mathbb{W}, \mathbf{u}) \|^2 \, d\mathbf{u} - 2 \int p(\mathbf{u}) \sum_{k=1}^{K} g_k(\mathbf{u}) \, \varphi_k(\mathbb{W},\mathbf{u}) \, d\mathbf{u} + \varepsilon_1,$$

où ε_1 ne dépend pas des φ_k. Minimiser $\varepsilon(\infty)$ revient donc à minimiser

$$\text{(A–3)} \qquad \varepsilon(\infty) - \varepsilon_2 = \int p(\mathbf{u}) \, \| \mathbf{g}(\mathbf{u}) - \varphi(\mathbb{W}, \mathbf{u}) \|^2 \, d\mathbf{u}.$$

L'application φ construite approxime donc $\mathbf{g}(\mathbf{u})$. Autrement dit, si Φ est un ensemble suffisamment large, trouver la sortie φ_k la plus élevée revient à trouver la sortie g_k (ou b_k) la plus élevée. ◊

♦Démonstration du théorème II

Ici, $y_i^{(n)} = \kappa(i,k)$ si $\mathbf{x}^{(n)} \in \omega_k$. D'après (A–1) nous avons donc

$$\text{(A–4)} \qquad \varepsilon(\infty) = \int p(\mathbf{u}) \, \| \varphi(\mathbb{W}, \mathbf{u}) \|^2 \, d\mathbf{u} - 2 \sum_{k=1}^{K} \sum_{i=1}^{K} \int p(\mathbf{u}/\omega_k) \, P_k \, \kappa(i,k) \, \varphi_i(\mathbb{W}, \mathbf{u}) \, d\mathbf{u} + \varepsilon_3,$$

où ε_3 ne dépend pas des φ_i. On pose cette fois $G_i(\mathbf{u}) = B_i(\mathbf{u})/p(\mathbf{u})$, où $B_i(\mathbf{u})$ est définie en (3–6). La relation (A–4) devient alors

$$\varepsilon(\infty) = \int p(\mathbf{u}) \| \varphi(\mathbb{W},\mathbf{u}) \|^2 \, d\mathbf{u} - 2 \sum_{i=1}^{K} \int p(\mathbf{u}) \, G_i(\mathbf{u}) \, \varphi_i(\mathbb{W},\mathbf{u}) \, d\mathbf{u} + \varepsilon_4.$$

Minimiser $\varepsilon(\infty)$ revient donc finalement à minimiser

$$\text{(A–5)} \qquad \varepsilon(\infty) - \varepsilon_5 = \int p(\mathbf{u}) \, \| G(\mathbf{u}) - \varphi(\mathbb{W},\mathbf{u}) \|^2 \, d\mathbf{u},$$

ce qui montre que $\varphi(\mathbb{W},\mathbf{u})$ approxime $G(\mathbf{u})$. Ici encore, si Φ est suffisamment large, pour la plupart des $\mathbf{u}$ fixés les applications $G(\mathbf{u})$ et $\varphi(W,\mathbf{u})$ seront minimales pour la même composante k. ◊

♦Lemme

Les théorèmes III et IV résultent d'un lemme commun. Ecrivons l'erreur quadratique (3–12) sous la forme

$$\varepsilon(N,Q) = \sum_{k=1}^{K} \frac{N_k}{N} \frac{1}{N_k} \sum_{\mathbf{x}^{(n)} \in \omega_k} \frac{1}{Q} \sum_{q=1}^{Q} \| \mathbf{y}^{(n)} - \varphi(\mathbb{W}, \mathbf{x}^{(n)} + \mathbf{z}^{(n,q)}) \|^2.$$

Posons à présent

$$\text{(A–6)} \qquad \hat{P}_k = \frac{N_k}{N},$$

$$\text{(A–7)} \qquad \xi_k(\mathbf{u}) = \| \mathbf{y}^{(n)} - \varphi(\mathbb{W}, \mathbf{u}) \|^2, \text{ pour } \mathbf{x}^{(n)} \in \omega_k,$$

ce qui est possible car la sortie $\mathbf{y}^{(n)}$ ne dépend que de k lorsque $\mathbf{x}^{(n)} \in \omega_k$. Quand Q tend vers l'infini, nous avons:

$$\varepsilon(N,\infty) = \sum_{k=1}^{K} \hat{P}_k \int p_z(\mathbf{u}) \frac{1}{N_k} \sum_{\mathbf{x}^{(n)} \in \omega_k} \xi_k(\mathbf{x}^{(n)} + \mathbf{u}) \, d\mathbf{u}.$$

Soit après le changement de variable $\mathbf{v} = \mathbf{x}^{(n)} + \mathbf{u}$, et en posant

$$\text{(A–8)} \qquad \hat{p}(\mathbf{v}/\omega_k) = \frac{1}{N_k} \sum_{\mathbf{x}^{(n)} \in \omega_k} p_z(\mathbf{v} - \mathbf{x}^{(n)}),$$

nous obtenons

$$\text{(A–9)} \qquad \varepsilon(N,\infty) = \sum_{k=1}^{K} \hat{P}_k \int \hat{p}(\mathbf{v}/\omega_k) \, \xi_k(\mathbf{v}) \, d\mathbf{v}. \qquad ◊$$

♦Démonstration du théorème III

Ici, $y_j^{(n)} = \delta_{jk}$ si $\mathbf{x}^{(n)} \in \omega_k$. On définit les quantités suivantes:

$$\text{(A–10)} \qquad \hat{p}(\mathbf{v}) = \sum_{k=1}^{K} \hat{P}_k \, \hat{p}(\mathbf{v}/\omega_k)$$

et

$$\text{(A–11)} \qquad \hat{g}_k(\mathbf{v}) = \hat{P}_k \, \hat{p}(\mathbf{v}/\omega_k) \, / \, \hat{p}(\mathbf{v}).$$

Alors l'expression (A–9) de l'erreur devient

$$\varepsilon(N,\infty) = \int \hat{p}(\mathbf{v}) \| \varphi(\mathbf{v}) \|^2 \, d\mathbf{v} - 2 \int \sum_{k=1}^{K} \hat{g}_k(\mathbf{v}) \, \hat{p}(\mathbf{v}) \, \varphi_k(\mathbb{W},\mathbf{v}) \, d\mathbf{v} + \varepsilon_1.$$

Et finalement

$$\text{(A–12)} \qquad \varepsilon(N,\infty) = \int \hat{p}(\mathbf{v}) \, \| \varphi(\mathbb{W},\mathbf{v}) - \hat{\mathbf{g}}(\mathbf{v}) \|^2 \, d\mathbf{v} + \varepsilon_2$$

montre que la fonction $\varphi(\mathbb{W},\mathbf{v})$ approxime la fonction $\hat{\mathbf{g}}(\mathbf{v})$, qui est elle-même une estimation de la fonction bayesienne $\mathbf{g}(\mathbf{v})=\mathbf{b}(\mathbf{v})/p(\mathbf{v})$ définie en (3–7). La conclusion est similaire à celle du théorème I. ◊

♦Démonstration du théorème IV

Le codage est défini maintenant par (3–9):

$$y_i^{(n)} = \kappa(i,k) \text{ pour } \mathbf{x}^{(n)} \in \omega_k.$$

On garde la notation (A–10) et on pose $\hat{G}_i(\mathbf{u}) = \hat{B}_i(\mathbf{u})/\hat{p}(\mathbf{u})$, avec

$$\text{(A–13)} \qquad \hat{B}_i(\mathbf{x}) = \sum_{j=1}^{K} \kappa(i,j) \, \hat{P}_j \, \hat{p}(\mathbf{x}/\omega_j).$$

Il vient alors d'après (A–9) que

$$\varepsilon(N,\infty) = \int \hat{p}(\mathbf{v}) \| \varphi(\mathbf{v}) \|^2 \, d\mathbf{v} - 2 \int \sum_{i=1}^{K} \hat{G}_i(\mathbf{v}) \, \hat{p}(\mathbf{v}) \, \varphi_i(\mathbb{W},\mathbf{v}) \, d\mathbf{v} + \varepsilon_3.$$

D'où finalement

$$\text{(A–14)} \qquad \varepsilon(N,\infty) = \int \hat{p}(\mathbf{v}) \, \| \varphi(\mathbb{W},\mathbf{v}) - \hat{G}(\mathbf{v}) \|^2 \, d\mathbf{v} + \varepsilon_4. \qquad ◊$$