REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

==========

UNIVERSITE MENTOURI – CONSTANTINE –

FACULTE DES SCIENCES EXACTES

==========

DEPARTEMENT DE MATHEMATIQUES

N° d'ordre : **110 / T.E./ 2010**

N° de série : **02 / MAT / 2010**

THESE PRESENTEE POUR L'OBTENTION

DU

DIPLOME DE DOCTEUR D'ETAT

EN

MATHEMATIQUES

« Théorie de l'information, complexité et réseaux de neurones »

Par

Ahmed CHIBAT

OPTION

STATISTIQUE APPLIQUEE

Devant le jury :

Mr M.N . BENKAFADAR	Prof.	Université Mentouri Constantine	Président
Mr F.L. RAHMANI	M.C.	Université Mentouri Constantine	Rapporteur
Mme S. BOUGHABA	M.C.	Université Mentouri Constantine	Examinateur
Mr M. BOUZIT	M.C.	Université Larbi Ben M 'Hidi O.E.B.	Examinateur
Mr A. NOUAR	M.C.	Université de Skikda	Examinateur
Mr M. REGHIOUA	M.C.	E.N.S. Constantine	Examinateur

Soutenue le 20 NOVEMBRE 2010

CHAPITRE INTRODUCTIF

Traditionnellement, il est attendu de voir l'introduction d'une thèse se développer autour de trois points importants :

Situation du contexte du travail
Plan de la thèse et exposition des objectifs
Présentation succincte des résultats obtenus.

Cependant, dans notre cas, le premier point déjà ne saurait se suffire d'une simple description. Il exige, à lui seul, de conduire tout un argumentaire. Il nous semble impératif, en effet, de concilier au préalable le domaine de la thèse, qui est la statistique appliquée, et son titre qui évoque la théorie de l'information et les réseaux de neurones.

Notre travail se situe bel et bien, et d'emblée, dans le cadre de la statistique appliquée mais il se fait par le biais de ces deux disciplines qui, bien que nées partiellement, l'une et l'autre, en dehors du fief traditionnel de la statistique, ont contribué fortement à apporter à cette science autant un nouvel éclairage dans les concepts et les fondements, que de précieuses innovations dans les outils et techniques.

D'un côté, tout un travail de réinterprétation de la statistique s'est fait à la lumière de la théorie de l'information. Cette dernière a stimulé prodigieusement la recherche en statistique sous la forme de vagues successives qui, depuis une soixantaine d'années, continuent de produire des résultats déterminants. Dans beaucoup d'aspects, l'empiètement de la théorie de l'information et de la science statistique est tel qu'elles en deviennent indissociables.

De l'autre côté, les réseaux de neurones qui ont pris leurs sources, entre autres, dans le domaine informatique, sont aujourd'hui assimilés simplement à une extension des méthodes statistiques. Plus qu'à reproduire les méthodes standards de la statistique, ils sont surtout aptes à apporter des solutions nouvelles et à juguler des problèmes traditionnellement classés difficiles.

Ce qui vient d'être avancé nécessite de notre part une justification claire et bien établie.

Nous nous devons donc de démontrer que l'essor de la recherche en statistique est grandement redevable à ces deux disciplines.

De quelle manière les problématiques qui relevaient de la statistique classique sont elles devenues des questions centrales dans ces nouvelles spécialités ?

De quelle manière ont-elles été traitées ?

Quel est l'apport de ces nouvelles approches ?

Quel est l'étendue de l'ouverture offerte pour la recherche en statistique ?

Au milieu de ces empiètements, enchevêtrements et entrelacements, où se situent désormais les frontières actuelles de la statistique, et qui doivent être bien loin des frontières traditionnellement reconnues?

Nous avons pensé que le meilleur moyen d'apporter des réponses édifiantes à toutes ces questions est de procéder à un survol de l'historique de chacune de ces deux disciplines.

C'est cela qui pourrait permettre d'observer la maturation des idées et des concepts, de noter les moments de leurs implications dans le traitement des considérations d'ordre statistique, d'évaluer la pertinence de leurs apports et d'entrevoir la ligne de front qui s'entrouvre dans la recherche grâce à leurs contributions.

Historique **de** la théorie de l'information

L'historique de la théorie de l'information que allons développer s'articule sur trois parties :

- La période précédant les travaux de Shannon sur « la théorie mathématique de la communication ».
- La décennie ayant suivi les travaux de Shannon, et qui a été la période déterminante dans l'implication de la théorie de l'information dans la science statistique.
- La décennie ultérieure qui a vu la théorie de l'information dépasser le cadre statistique, stimuler la recherche, et provoquer l'émergence de nouvelles disciplines mathématiques.

La présentation thématique requise par la thèse nous invite à axer notre présentation sur la statistique en premier lieu, puis sur les mathématiques en général. Pour cette raison, c'est la deuxième partie de l'historique qui exige le plus d'attention puis, à un moindre degré, la troisième partie.

Notons, cependant, que les concepts de la théorie de l'information, de même que son formalisme, ont mis du temps à se concevoir. Les éléments relevant de la statistique sont apparus dès les premières heures. Les idées sont venues à maturité graduellement, dans plusieurs disciplines, pour donner finalement une théorie mathématique et unifiée. C'est cela qui a justifié l'existence de la première partie de l'historique.

Notons aussi, que l'expansion de la théorie de l'information a largement transcendé les frontières des disciplines qui nous intéressent. Nous allons omettre ces développements car ils sortent du cadre de notre recherche, néanmoins nous n'allons pas manquer de dire que la statistique a encore profité,

soit directement soit indirectement, de l'investissement de la théorie de l'information de ces domaines divers et apparemment lointains.

Première partie : émergence de la théorie de l'information

L'information, nouvelle notion scientifique, apparaît progressivement dans différents domaines du savoir avant de devenir un véritable concept doté d'un formalisme et de faire l'objet, dans les années 40 d'une théorie.

Elle est apparue dès 1922 dans trois disciplines différentes :

Statistique : R.A. Fisher dès 1922 [1]

L'information est fondée sur la notion de vraisemblance et placée dans le cadre de la théorie de l'estimation statistique. Le concept naît de préoccupations méthodologiques relatives à la conception des plans d'expériences.

Télécommunications : Harry Nyquist en 1924 [3]

C'est le terme '*intelligence'* qui est donné pour désigner ce que ultérieurement sera appelé *information*. Le concept est technique et il concerne la mesure de l'efficience de différents systèmes de télécommunications.

Physique : G.N. Lewis en 1930.[3]

La notion d'information se dégage lors du traitement de la question relative à la nature de l'entropie et du second principe de la thermodynamique.

Travaux de Fisher

En 1922, dans son article : "*On the Mathematical Foundations of Theoretical Statistics",* publié dans *Philosophical Transactions of the Royal Society*, **R.A.** Fisher [1] est probablement le premier à chercher un fondement à l'emploi de la notion 'd'information', en la plaçant dans le cadre de sa théorie de l'estimation statistique, fondée sur la notion de vraisemblance.

Il introduit dans ses écrits la notion scientifique d'information de cette manière :

> "(…) l'objet de la méthode statistique est la réduction des données. Une masse de données, si importante qu'elle en est inintelligible, doit être remplacée par un relativement petit nombre de quantités qui doivent représenter correctement cette masse, ou, en d'autres mots, doivent contenir la plus grande part possible, si ce n'est la totalité de **l'information pertinente contenue dans les données** d'origine."

Ici, Fisher sous-entend une information d'ordre sémantique, on peut comprendre le mot *'information'* dans le sens du mot *'renseignement'*. Il n'y a pas d'indication apparente concernant les variables aléatoires et ce concept ne peut pas, à ce niveau, être déjà lié à l'information de Shannon.

Cette évocation de l'information est faite de manière qualitative mais, jusqu'en 1935, Fisher [4, 5, 6, 7, 8, 9, 10] va préciser cette notion pour parvenir, au fil de ses publications, à une

définition mathématique. Il montre qu'une statistique est une transformation des données de l'échantillon qui le résume mais aussi le simplifie. Une statistique est dite exhaustive si le résumé qu'elle constitue ne supprime rien de 'l'information' (dans le sens de 'renseignement') qui est contenue dans l'échantillon.
Dans le cadre de sa théorie de l'estimation, la définition de l'information, au sens de Fisher, s'inscrit au sein d'une structure statistique paramétrée et concerne avant tout ce que l'échantillon peut enseigner à propos de la valeur du paramètre. Le problème que se pose Fisher {1] dans sa publication de 1922 relève de la théorie de l'estimation : le problème qu'il se pose est de pouvoir 'estimer', à partir des échantillons relevés, les valeurs des paramètres caractéristiques des distributions de probabilité d'une population hypothétique (par exemple pour le calcul des paramètres *m* et σ d'une loi normale). Il propose dans ce cadre un « traitement quantitatif de l'information apportée par un échantillon », qui permet déjà des applications dans le choix des courbes d'erreurs qui permettent de considérer l'échantillon comme étant le plus représentatif possible de la population envisagée.

C'est dans la publication de 1925 consacrée à la " Théorie de l'estimation statistique ", que les premières définitions quantitatives apparaissent, et Fisher [5] donne l'expression générale de la quantité d'information relative à un paramètre θ apportée par un échantillon de n observations :

$$S\left\{\frac{1}{m}\left(\frac{\partial n}{\partial \theta}\right)^2\right\}$$

où S indique une sommation pour toutes les observations et m le produit de n par la probabilité pour une observation de tomber dans l'une des classes préalablement déterminées

Cela c'est fait à travers le traitement d'un exemple de calcul d'une quantité d'information dans le cas de deux estimations différentes du paramètre σ d'une loi normale, à partir d'un même échantillon de n valeurs observées.

En 1928, dans la deuxième édition des *Méthodes statistiques pour les chercheurs*, Fisher redéfinie mathématiquement à nouveau l'information et la pose égale à l'inverse de la variance dans le cas d'une distribution normale des valeurs d'un paramètre estimé.

En fait, l'information qui est au départ " contenue dans les données " devient peu à peu la caractéristique d'une statistique. L'information se voit donc définie comme une grandeur constante, calculée à partir d'un échantillon.

Dans le cas de plusieurs paramètres, la " matrice information" est définie par

$$F_{ij} = \sum_k \frac{1}{m_k}\left(\frac{\partial n_k}{\partial \varphi_i}\frac{\partial n_k}{\partial \varphi_j}\right)$$

où k représente le nombre de classes dans lesquelles on range les observations, i et j variant entre 1 et le nombre de paramètres.

Prolongements des travaux de Fisher

En partant des travaux de Fisher, la notion scientifique d'information connaît d'importants développements. Principalement grâce aux travaux de J.L. Doob (1936, 1941) et de A. Bhattacharyya (en trois parties de 1946 à 1948).

Partant de l'axiomatisation des probabilités proposée par Kolmogorov en 1933, et donc de la théorie de la mesure, J.L. Doob intègre la théorie du maximum de vraisemblance de Fisher, ainsi que la définition de l'information qui en découle, dans le cadre de la théorie des probabilités, et, plus précisément, dans le cadre de la théorie mathématique des ensembles mesurables, fondée par les travaux de Borel et Lebesgue.

En 1936, dans sa publication sur " *l'estimation statistique* ", J.L. Doob [12] reprend certains résultats de Fisher avec plus de rigueur mathématique et dans son article de 1941 [13] il définit la " *probabilité comme une mesure* ", s'opposant à l'approche fréquentielle.

Les définitions suivantes ont vu le jour :

la quantité d'information est

$$I(\theta) = E_\theta \left[\left(\frac{\partial}{\partial \theta} \log p(x, \theta) \right)^2 \right]$$

Où E_θ représentante l'espérance mathématique

De même, dans le cas d'un modèle dominé (selon la décomposition de Lebesgue et de Radon-Nikodym) et d'une vraisemblance L associée à un paramètre θ, on trouve alors, la définition suivante

$$I_{ij}(\theta) = E_\theta \left[\frac{\partial}{\partial \theta_i} \log L(\theta) \frac{\partial}{\partial \theta_j} \log L(\theta) \right]$$

Entre 1946 et 1948, A. Bhattacharyya [14] publie dans Sankhya, la revue indienne de statistiques, un long article en trois parties " Sur des équivalents de la quantité d'information et leur utilisation en statistique ". Travail qui vise à améliorer l'expression de la quantité d'information dans la théorie de l'estimation de Fisher.

1948 et la théorie mathématique de la communication

La notion d'information est présentée par C. Shannon, dans le cadre d'une théorie mathématique, dans une plus grande abstraction qui rend accessoire la référence à un système physique concret.

Shannon se présente dans la droite lignée des travaux effectués dans le domaine des télécommunications, principalement par H. Nyquist (1924) et R.V.L.Hartley (1928).

Dans son article de 1924, Nyquist [2] donne une expression quantitative concernant la vitesse à laquelle on peut transmettre l'information (*'intelligence'*). Les deux facteurs qu'il étudie sont " le choix des codes " et " la mise en forme du signal (*'signal shaping'*) ".

Il est constaté que la quantification de la notion d'information s'accompagne d'un renoncement à la modélisation de la dimension sémantique de l'information.

Nyquist donne dans le corps de son texte pour la " *vitesse de transmission de l'intelligence* " la formule :

$$W = Klog\ m$$

où K est une constante et m est le nombre de valeurs distinguées du courant.

La publication de R.V.L. Hartley [15] de 1928, " *Transmission d'information* ", est souvent mentionnée comme une des origines de la théorie mathématique de la communication.

Le référent technique de Hartley est le système Baudot, dont le fonctionnement était caractérisé par un code de longueur constante, cinq symboles primaires, quelle que soit la lettre à transmettre. A partir de là il obtient une formule donnant la quantité d'information,

$$H = nlog\ s$$

où n représente le nombre de sélections et s le nombre de symboles primaires disponibles

Cette formule concerne toutes les combinaisons physiquement possibles, indépendamment du fait qu'elles peuvent ne pas correspondre à un des arrangements de symboles primaires utilisés dans le code. Il s'agit donc d'une mesure '*physique*' d'une quantité que l'on pourrait qualifier de '*virtuelle*' puisqu'elle concerne ce qu'on *pourrait* transmettre, au conditionnel, et pas ce qu'on transmet effectivement. Cette remarque a ici son importance car elle reste valable tout le long du développement de ce qu'il est convenu d'appeler depuis la fin des années 40 la '*théorie de l'information*'.

Les travaux de Claude E. Shannon

C'est sur une période d'au moins dix ans, entre 1939 et 1948, que naît la théorie mathématique de la communication dont Shannon est l'un des principaux auteurs.

Claude E. Shannon [16] s'était proposé l'étude du point de vue mathématique des problèmes qui concernent la transmission à l'aide de signaux dans les systèmes de communications électriques. De cette étude, le saut qualitatif dû à Claude E. Shannon vient de ce qu'il a réussi à s'élever par le mode de traitement de ces problèmes particuliers, à un très grand degré de généralité, créant ainsi la théorie mathématique de l'information. Claude E. Shannon a donné une mesure de la quantité de l'information qui se transmet à l'aide de signaux, a étudié du point de vue mathématique la transmission de l'information par des voies et a démontré l'existence de possibilité pour réduire l'influence des perturbations qui altèrent l'information transmise. Les résultats qu'il a obtenus ont dépassé par leur généralité les systèmes de communication particuliers et ont permis la description des phénomènes liés à la transmission de l'information dans tout système de communication.

C'est en 1948 que paraît, en deux parties, l'article du *Bell System Technical Journal* précisément intitulé " *A Mathematical Theory of Communication* ".

La publication de Shannon concerne la transmission d'information, en distinguant d'une part le caractère continu ou discret de la source d'information, d'autre part la présence ou l'absence de bruit dans la voie de communication.

la théorie de Shannon se présente comme une théorie mathématique, sous la forme d'une suite de 23 théorèmes.

Shannon introduit la grandeur

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

où p_i sont les probabilités de sélection.

Shannon montre qu'en partant de quelques propriétés requises comme celle d'additivité et de continuité, cette expression est la seule qui puisse convenir, la constante *K* étant quelconque, déterminant l'unité d'information. Il appelle cette expression '*entropie'*.

Il définit aussi la quantité d'information d'une source continue avec la grandeur

$$\int_{-\infty}^{+\infty} p(x) \log p(x) dx$$

où $p(x)$ représente une densité de probabilité.

La théorie de Shannon constitue une première '*axiomatisation*' de la notion d'information.

Rapidement la théorie exposée par Shannon se nommera d'ailleurs théorie" de l'information " et non plus "de la communication ".

Répercussions des travaux de Shannon :

L'œuvre de Shannon ne constitue pas seulement le traitement de questions théoriques relevant de l'engineering mais s'élève comme une œuvre mathématique. C'est au vu de cela que des mathématiciens des plus éminents se sont intéressés à cette théorie.

Les liens entre les différentes définitions de l'information, surtout celles de Fisher et de Shannon, ont fait l'objet de premières recherches, au début des années 50, qui ont généralement précédé le recueil d'articles de Khinchin [17] paru aux Etats-Unis en 1957 portant sur l'étude des *Fondements mathématiques de la théorie de l'information*.

Dans un troisième temps, au milieu des années 60, on assiste à partir de ces travaux à la naissance d'une théorie mathématique de la complexité qui amènera des mathématiciens comme Kolmogorov [18, 19] à repenser le calcul des probabilités à partir de la notion d'information et un ingénieur d'I.B.M., Gregory Chaitin [20], à fonder la théorie 'algorithmique' de l'information.

2ème partie : théorie de l'information et statistique

Schützenberger et Mandelbrot

Dès 1949 , Schützenberger [21] tente de rapprocher la théorie de la communication de la théorie des jeux, pour laquelle von Neumann et Morgenstein venaient de proposer un nouveau formalisme.

C'est en 1949 aussi que Schützenberger reçoit de R. A. Fisher une carte de félicitations pour un audacieux rapprochement entre la théorie des plans d'expériences et celle des codes correcteurs d'erreurs

Dans sa thèse, parue en 1953, Schützenberger [22] propose une " *mesure quelconque de la quantité d'information* ", $H(x) = xDlog\ x + (1-x)\,\mathrm{Dlog}(1-x)$, où D est un opérateur linéaire quelconque, pour le cas d'une variable aléatoire binaire.

L'information de Fisher se retrouve avec $D = (d^2/d\theta^2)$

et celle de Shannon pour $D = -1/log_2$.

En 1952, Mandelbrot [23] introduit une forme générale pour l'expression de la quantité d'information $\mathrm{H} = \sum x\mathrm{S}(\log x)$ où S est un opérateur linéaire quelconque qui permet de retrouver la formule de Shannon pour $\mathrm{S} = \mathrm{C^{te}}$ et celle de Fisher pour $\mathrm{S} = \partial^2/\partial\theta^2$.

Kullback et Leibler

En 1951, Solomon Kullback et R.A. Leibler[24], publient un article dans les Annals of Mathematical Statistics intitulé " *Sur l'information et l'exhaustivité* ['*sufficiency*']".

pour Kullback [25], cette publication sera l'amorce de tout un travail de réinterprétation des statistiques autour de la notion d'information, comme en témoigne son livre, *Théorie de l'information et statistique*, paru pour la première fois en 1959.

Des travaux de Kullback, retenons la définition, introduite dès 1951 [24], de l'information moyenne apportée par un échantillon en faveur d'une hypothèse H_1 contre une hypothèse H_2 et surtout celle de la notion de divergence :

Si μ_1 et μ_2 sont deux mesures de probabilités absolument continues l'une par rapport à l'autre associées aux deux hypothèses et λ une autre mesure également absolument continue par rapport *aux* deux autres, le théorème de Radon-Nikodym permet d'affirmer qu'il existe deux fonctions f_1 et f_2, uniques en dehors d'ensembles de mesure nulle pour λ et appelés 'densités de probabilités généralisées', telles que

$$\mu_i(E) = \int_E f_i(x)d\lambda(x), \qquad i = 1{,}2$$

Dans ce cas, l'information apportée par $\mathrm{X} = x$ pour discriminer H_1 *de* H_2 est définie par $\log\frac{f_1(x)}{f_2(x)}$.

L'information moyenne est alors donnée par

$$I(1:2) = \int f_1(x)\log\frac{f_1(x)}{f_2(x)}d\lambda(x)$$

et la divergence n'est autre que $J(1,2) = I(1:2) + I(2:1)$.

La '*divergence'* introduite par Kullback et Leibler est également appelée *'entropie croisée', 'écart entropique' ou 'information de Kullback-Leibler',* additive pour des observations indépendantes et toujours positive, nulle si les lois sont identiques.

Grâce à cette définition de l'information, on ne tente plus de définir l'information de façon absolue mais au contraire de façon relative en prenant en compte deux distributions de probabilités. La disparité entre les définitions proposées par Shannon dans les cas discrets et continus est ainsi levée.

Kullback, dans l'introduction de son livre paru en 1959 [25], situe cette théorie dès le début comme une " *branche de la théorie mathématique des probabilités et des statistiques* ".

La théorie de l'information doit rester une théorie mathématique et Kullback montre que celle-ci est à replacer dans un cadre plus général que celui qui a servi à Shannon pour l'élaboration de sa théorie de la communication. Ainsi note-t-il, dans sa publication de 1951, que la définition de l'information proposée par Shannon n'est qu'un cas particulier de la théorie plus générale qu'il propose.

Kullback et Leibler replacent les deux définitions de la quantité d'information dans un cadre mathématique plus général inaugurant une théorie de l'information " *unifiée* ".

La statistique inférentielle se trouve au centre de ses préoccupations puisque c'est de là qu'il reprend l'idée selon laquelle toute variable observée ou mesurée peut être modélisée, d'un point de vue mathématique, par une variable aléatoire soumise à une loi de probabilité. Il s'agit alors depuis les premières publications de Fisher '*d'inférer'* la valeur de ces paramètres à partir de 'l'information' fournie par les échantillons.

Pour Kullback c'est en effet la théorie de l'information qui suscite des recherches en statistique visant à montrer que les travaux de Fisher ou Shannon peuvent être considérés comme des cas particuliers.

Jaynes

1957 : Jaynes [26] pose le principe de l'entropie maximale qui consiste à décider que lorsqu'on dispose d'informations partielles quant à l'issue d'un processus il faut choisir les probabilités de façon à maximiser l'incertitude concernant l'information manquante.

Le principe d'entropie maximale donnera lieu à la création d'une association scientifique d'envergure internationale qui organise chaque année des conférences sur le thème " *Maximum Entropy and Bayesian Methods* ", l'accent étant porté depuis le milieu des années 70 sur l'analyse de données.

De plus, il a été démontré par la suite que le principe introduit par Jaynes s'identifie à celui que l'on trouve dans le livre de Kullback et qu'on nomme, par symétrie, le '*principe de divergence minimale'*, en référence à la notion introduite par Kullback et Leibler en 1951.

Aux Etats-Unis déjà, à la fin des années 50, on estime, en référence à tous ces travaux ainsi qu'à ceux de A. Feinstein [27], MacMillan [28] et Woodward [29], que les bases mathématiques de la théorie de l'information sont clairement établies.

Troisième partie : complexité

Khinchin et Kolmogorov

En 1957 paraît un livre sur les fondements de la théorie de l'information rassemblant deux articles de 1953 et 1956 du mathématicien soviétique A.I. Khinchin [17].

Il ne s'agit cependant pas tant de 'fondements' permettant de construire une théorie logiquement consistante, mais davantage de s'intéresser aux bases mathématiques de la théorie de l'information.

Notant que Shannon ne donne pas de démonstration satisfaisante sur le plan mathématique de son théorème fondamental, il présente rapidement les principaux résultats de l'ingénieur mathématicien américain et fournit la première preuve rigoureuse de ce qui correspond dans la publication de Shannon aux théorèmes 3 et 4 de la voie sans bruit ainsi qu'à l'annexe 3 sur les sources ergodiques.

Khinchin et Kolmogorov, tous deux lauréats du Prix Staline (en 1941) ayant commencé à travailler sur les probabilités dès les années 20, sont dans les années 50 parmi les principaux mathématiciens à travailler autour de la notion d'information.

Au début des années 50, Kolmogorov [30] voit dans la théorie de l'information le moyen de décrire de façon universelle un objet mathématique :

si un 'objet aléatoire' ξ prend les valeurs $x_{1...}x_n$ avec les probabilités $p_{1...}p_n$, la quantité définie par $H(\xi) = -\Sigma\, p_k \log p_k$, l'entropie, représente le nombre de bits qui suffisent en moyenne à décrire ξ.

Kolmogorov dirige les thèses d'une soixantaine d'étudiants et crée une véritable 'école soviétique de la théorie de *l'information'*. Dans les années qui suivent, il coordonne toute une série des recherches tournant autour de la théorie de l'information et cette orientation se traduit dans le choix des articles paraissant dans la revue intitulée *Théorie des probabilités et applications*. Dès 1959, on trouve ainsi un article de Y.V. Linnik démontrant une version du théorème central limite à partir de la théorie de l'information.

A partir de 1965 Kolmogorov participe à la fondation de la théorie de la complexité algorithmique.

Dans sa publication de 1965, Kolmogorov [18] fait état de " Trois approches de la définition quantitative de l'information'' :

- l'approche combinatoire : si une variable x prend ses valeurs dans un ensemble X contenant N éléments, on peut commencer par définir l'information par $\mathrm{H(X)} = \log_2 \mathrm{N}$. l'intérêt de cette démarche est de montrer qu'il est possible de définir quantitativement l'information sans introduire les probabilités
- l'approche probabiliste .

- l'approche " *algorithmique* " : l'information peut être définie au niveau individuel, et non en référence à toute une collection de messages comme dans la théorie probabiliste de Shannon qui repose, en définitive, sur la théorie de la mesure. L'information apportée par une série de nombres sur une autre est mesurée à une constante additive près en fonction de l'algorithme utilisé pour construire la seconde série.

Kolmogorov définit la complexité K(x) en fonction de la taille minimale d'un tel programme.

Dans le cadre d'une analyse plus générale de la notion de hasard, il affirme que l'approche algorithmique permet non seulement de définir l'information indépendamment des probabilités (comme dans l'approche combinatoire) mais aussi d'entrevoir un moyen de " fournir une nouvelle base au calcul des probabilités ".

Solomonoff, Martin-Löf et Chaitin

Dans l'élaboration de ce qu'on allait bientôt appeler 'la théorie algorithmique de l'information' ou 'la théorie des probabilités algorithmiques', on trouve quatre scientifiques qui, plus ou moins indépendamment les uns des autres, sont parvenus dans les années 60 à des résultats comparables. A côté de Kolmogorov, il y a eu Ray J. Solomonoff, Martin-Löf et Chaitin.

1964 : Ray J. Solomonoff [31] démontre le " théorème fondamental" de la complexité algorithmique.
1964 : Martin-Löf démontre à son tour, indépendamment de Solomonoff, le théorème fondamental qui assure qu'il existe un algorithme de référence permettant de définir la complexité d'une suite de nombres.
1974 : Chaitin [20] crée : la " théorie algorithmique de l'information ".

Les études sur la complexité de Kolmogorov ou la théorie algorithmique de l'information de Chaitin (selon qu'on se réfère à la dimension mathématique ou à l'importance de la notion d'information définie entre autres par Chaitin) constitue aujourd'hui un des domaines de recherche les plus vivants concernant la notion scientifique d'information.

Historique réseaux de neurones

Il est admis que les travaux sur les réseaux de neurones artificiels ont commencés avec la publication de McCulloch et Pitts [42] en 1943. McCulloch était un psychiatre et un neuro anatomiste ; Pitts était un jeune mathématicien autodidacte.

Dans leur article, McCulloch et Pitts ont développé un calcul logique pour les réseaux de neurones qui unifie la neurophysiologie et la logique mathématique. Leur modèle formel suit la loi du 'tout ou rien'. Avec un nombre suffisant de ces unités simples, et des connections établies convenablement et fonctionnant d'une manière synchrone, McCulloch et Pitts ont montré qu'un réseau ainsi constitué peut, en principe calculer toute fonction calculable.

Le développement majeur qui a suivi a été le livre « *The Organization of Behavior* » publié par Hebb [43] en 1949, dans lequel une hypothèse explicite a été présentée pour la première fois, concernant la règle d'apprentissage physiologique se basant sur les modifications synaptiques. Hebb a proposé comme hypothèse que la connectivité du cerveau change continuellement lorsque l'organisme apprend les différentes tâches fonctionnelles, et des assemblages de neurones sont créés par de tels changements.

Le livre de Hebb a été une source d'inspiration pour le développement de systèmes apprenants et adaptatifs.

L'article de Rochester, Holland, Haibt et Duda (1956) [44], a peut être été la première tentative d'utilisation de la simulation par machine pour tester la théorie neuronale basée sur le postulat d'apprentissage de Hebb. Ces auteurs ont montré que *l'inhibition* doit être ajoutée pour que la théorie puisse fonctionner correctement.

Dans la même année (1956), Uttley [45] démontre qu'un réseau de neurones à synapses modifiables peut apprendre à classer des ensembles simples d'éléments binaires dans des classes correspondantes.

En 1979, Uttley [46] a émis l'hypothèse que l'efficacité d'une synapse variable dans le système nerveux dépend des relations statistiques entre les états fluctuants sur les deux côtés de la synapse. De cette manière, Uttley établit la relation avec la théorie de l'information.

En 1952, apparaît le livre d'Ashby [47], dans lequel il est stipulé que le comportement adaptatif n'est pas inné mais appris, et c'est à travers l'apprentissage que l'animal (système) évolue vers le meilleur.

Le livre met en exergue d'un côté, les aspects dynamiques des organismes vivants en qualité de machines et de l'autre côté le concept de stabilité.

En 1954, Minsky [48] rédige une thèse sur les réseaux de neurones et en 1961 [49] publie un article portant sur l'intelligence artificielle et contenant une longue section portant sur ce qui sera appelé 'réseaux de neurones'.

En 1967, Minsky [50] publie un livre qui étend les résultats de McCulloch et Pitts et les situe dans le contexte de la théorie des automates et de la théorie du calcul.

En 1954 également l'idée de filtre adaptatif non linéaire a été proposée par Gabor [51]. L'apprentissage s'accomplit en fournissant à la machine des échantillons d'un processus stochastique en même temps que la fonction cible que la machine est sensée produire.
En 1956, Taylor [52] initie un travail sur les *mémoires associatives*.
1961 a vu l'introduction de la *matrice d'apprentissage* par Steinbuch [53].

En 1972, Anderson [54], Kohonen [55] et Nakano [56] ont indépendamment introduit l'idée d'une mémoire à matrice de corrélation basée sur la règle d'apprentissage du produit sortant.

En 1956, von Neumann [57], en utilisant l'idée de la redondance, résout le problème de la construction d'un réseau fiable avec des neurones qui peuvent être des composants non fiables.

C'est ce qui a motivé Winograd et Cowan (1963) [58] a suggérer une représentation redondante distribuée pour les réseaux de neurones. Winograd et Cowan ont montré comment un grand nombre d'éléments peuvent représenter collectivement un concept unique, avec une amélioration dans la robustesse et le parallélisme.

En 1958, Rosenblatt [59] introduit une nouvelle approche de la reconnaissance des formes avec son travail sur le *perceptron*, une nouvelle méthode d'apprentissage supervisé. Il démontre l'important théorème de la *convergence du perceptron*.

En 1960, Widrow et Hoff introduisent l'algorithme des moindres carrés (LMS) et l'utilisent pour formuler *l'Adaline* (élément linéaire adaptatif). La différence entre le perceptron et l'Adaline réside dans la procédure d'apprentissage.

En 1962, Widrow [60] propose l'un des premiers réseaux adaptatifs entraînable à couches, le Madaline.

En 1967, Amari [61] utilise la méthode du gradient stochastique pour la classification de formes adaptative.

En 1969, Minsky et Papert [62] démontrent mathématiquement qu'il existe des limites fondamentales aux possibilités de calcul des perceptrons à une seule couche. Dans une brève section sur les perceptrons multicouche, ils ont statué qu'il n'y a pas de raisons de croire qu'une quelconque des limitations du perceptron à une seule couche puisse être surmontée dans le cas de la version multicouche.

Un problème important rencontré dans la conception d'un perceptron multicouche est celui de l'allocation de crédits (*credit assignement problem*), qui est celui d'allouer des crédits aux neurones cachés.

A la fin des années 60, la plupart des idées et des concepts nécessaires pour résoudre le problème de l'allocation de crédits ont été formulés ainsi que plusieurs des idées sous jacentes aux réseaux

récurrents qui ont depuis été connu sous le nom de réseaux de Hopfield. Cependant il a fallut attendre les années 80 pour qu'émerge les solutions de ces problèmes basiques. Il y a eu trois raisons qui ont provoqué ce moratoire de plus de dix ans :

- La première raison est technologique – il n'y avait pas d'ordinateurs pour l'expérimentation.
- La deuxième raison est financière – les financements se sont taris à la suite de l'article de Minsky et Papert.
- l'analogie des réseaux de neurones avec le modèle des verres de spin était prématurée, celui-ci n'a été inventé qu'en 1975 par Sherrington et Kirkpatrick.

Ces facteurs ont contribué d'une manière ou d'une autre à amoindrir l'intérêt dans les réseaux de neurones.

Une activité importante qui a émergée dans les années 70 est celle concernant les cartes auto-organisatrices utilisant l'apprentissage compétitif.

En 1973, le travail de simulation par machine de von der Malsburg [63] a peut être été le premier à démontrer l'auto-organisation.
En 1976, motivé par les cartes topologiquement ordonnées du cerveau, Willshaw et von der Malsburg [64] ont publié le premier article sur la formation de cartes auto-organisatrices.

Dans les années 80, des contributions majeures, à la théorie et à la conception des réseaux de neurones, ont été faites sur plusieurs fronts, et avec elles s'est produite une résurgence de l'intérêt pour les réseaux de neurones.

En 1980, Grossberg [65], construit, sur la base de ses précédents travaux (1972 [66], 1976a [67] et 1976b [68]), un nouveau principe d'auto-organisation appelé *théorie de la résonance adaptative* (ART).
En 1982, Hopfield [70] utilisa l'idée d'une fonction d'énergie pour formuler d'une nouvelle manière la compréhension du calcul fait par des réseaux récurrents ayant des connections synaptiques symétriques. Plus encore, il établit l'isomorphisme entre de tels réseaux récurrents et le modèle Ising utilisé en physique statistique.
En 1983, Cohen et Grossberg [70] établissent un principe général pour évaluer la stabilité de leur modèle (content-addressable memory) qui inclus, comme cas particulier, la version en temps continu du réseau de Hopfield.

Un autre développement important en 1982 a été la publication de l'article de Kohonen [71] sur les cartes auto-organisatrices utilisant des structures en treillis à une ou deux dimensions.

En 1985, Ackley, Hinton et Sejnowski [72] développent une machine stochastique appelée machine de Boltzmann, et qui a été la première réalisation réussie d'un réseau de neurones multicouche.

Bien que l'algorithme d'apprentissage de la machine de Boltzmann soit très lent, il a brisé le blocage psychologique en montrant que les spéculations de Minsky et Papert sont non fondées.

En 1986, Rumelhart, Hinton et Williams [73, 74] développent l'algorithme de rétro-propagation (*Back-propagation algorithm*) et publient le célèbre ouvrage en deux volumes, '*Parallel Dsitributed Processing : Explorations in the Microstructures of Cognition*' [75] .

Ce livre a eu une influence majeur sur l'utilisation de l'apprentissage par rétro propagation, qui a émergé comme l'algorithme d'apprentissage le plus populaire utilisé pour entraîner les perceptrons multicouche.

Dès cette découverte, il est devenu possible de réaliser une fonction non linéaire d'entrée/sortie sur un réseau en décomposant cette fonction en une suite d'étapes linéairement séparables. De nos jours, les réseaux multicouches et la rétro-propagation du gradient reste le modèle le plus étudié et le plus productif au niveau des applications.

En 1988, Linsker [76] utilisa des concepts abstraits provenant de la théorie de l'information pour formuler le *principe de l'information mutuelle maximum* (Infomax) concernant l'auto-organisation dans les réseaux perceptuels. Cet article a déclenché l'exploration d'autre modèles de la théorie de l'information pour résoudre une grande classe de problèmes connu sous le nom de '*déconvolution aveugle'*.

En 1988, Broomhead et Lowe [77] décrivent une procédure pour la conception de réseaux multicouche, à propagation avant utilisant les fonctions radiales de base (RBF), qui sont une alternative aux perceptrons multicouche.

L'article de Broomhead et Lowe a engendré beaucoup de recherches pour relier la conception des réseaux de neurones à un important domaine de l'analyse numérique et aussi des filtres linéaires adaptatifs.

En 1990, Poggio et Girosi [78] ont enrichi la théorie des réseaux RBF en y appliquant la théorie de régularisation de Tikhonov.

Au début des années 90, Vapnik [79, 80, 81] et ses coauteurs inventent des réseaux à apprentissage supervisé à grande puissance de calcul, appelés machines à vecteurs supports (SVM). Ces réseaux servent à résoudre des problèmes relevant de la reconnaissance des formes, la régression, l'estimation de densité.

Cette nouvelle méthode est basée sur les résultats de la théorie de l'apprentissage avec des échantillons de tailles finis.

Ces auteurs ont conçu ce qui est appelée la dimension VC (Vapnik, Chervonenkis) qui fournit une mesure de la capacité d'un réseau de neurones à apprendre d'un ensemble d'exemples.

CHAPITRE II

ELÉMENTS DE LA THÉORIE DE L'INFORMATION

Incertitude et information

Expérience aléatoire en tant qu'espace probabilisé

Soit une expérience A dont les résultats possibles sont un nombre fini d'événements élémentaires $a_1, \dots, a_n$. Posons $p_1, \dots, p_n$ les probabilités de réalisation de ces événements. Nous avons

$$p_k \geq 0, \qquad 1 \leq k \leq n, \qquad \sum_{k=1}^{n} p_k = 1$$

Cette expérience se traduit par un espace probabilisé $\{A, a, p(a)\}$.

Définition 1 : (de C. Shannon)

La mesure H de l'incertitude sur l'expérience A est donnée par :

$$H(p_1, \dots, p_n) = -\sum_{k=1}^{n} p_k \, log \, p_k \qquad (1)$$

Où le logarithme est pris dans la base 2.

Remarque

Le choix de la base 2 n'est pas essentiel, nous pouvons toujours passer d'une base à l'autre grâce à la formule

$$log_b m = log_b a \, log_a m$$

Définition 2

La mesure de l'incertitude donnée par la formule (1) s'appelle *l'entropie de l'expérience A.*

Propriétés de l'entropie

Propriété 1

$$H(p_1, \dots, p_n) \geq 0$$

Propriété 2

Si pour un certain indice *i* nous avons $p_i = 1$, alors

$$H(p_1, \dots, p_n) = 0$$

Propriété 3

Pour toute répartition $p_1, \dots, p_n$ nous avons

$$H(p_1, \dots, p_n) \leq H\left(\frac{1}{n}, \dots, \frac{1}{n}\right)$$

Propriété 4

$$H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$$

Propriété 5

Soit A et B deux expériences indépendantes. Nous avons l'égalité

$$H(A, B) = H(A) + H(B)$$

Soit A une expérience quelconque pouvant produire l'un des n événements $a_1, \dots, a_n$. Soit B une autre expérience mettant en évidence *m* événements $b_1, \dots, b_m$ dont la réalisation est directement influencée par le résultat de l'expérience A.

Pour chaque événement b_l nous avons *n* probabilités $q_{1l}, \dots, q_{nl}$ où $q_{kl} \geq 0$, $1 \leq k \leq n$, $1 \leq l \leq m$, avec $\sum_{l=1}^{m} q_{kl} = 1$, $1 \leq k \leq n$.
q_{kl} représente la probabilité de réalisation de l'événement $b_l$dans l'expérience B en supposant que l'expérience A nous ait fourni l'événement a_k.

Définition 3

La mesure de l'incertitude donnée par

$$H_k(B) = H(q_{k1}, \dots, q_{km}) = -\sum_{l=1}^{m} q_{kl} \, log \, q_{kl}$$

s'appelle *l'entropie conditionnelle de l'expérience B quand l'événement a_k est donné*.

Définition 4

La mesure de l'incertitude donnée par

$$H_A(B) = \sum_{k=1}^{n} p_k \, H_k(B)$$

s'appelle *l'entropie conditionnelle de l'expérience B quand l'expérience A est donnée*.

Propriété 6

Soit A et B deux expériences quelconques. Nous avons l'égalité

$$H(A,B) = H(A) + H_A(B)$$

Propriété 7

Soit A et B deux expériences quelconques. Nous avons

$$H_A(B) \leq H(B)$$

L'égalité n'étant possible que si, et seulement si, les deux expériences sont indépendantes.

Propriété 8

Soit A et B deux expériences quelconques. Nous avons l'inégalité

$$H(A,B) \leq H(A) + H(B)$$

L'égalité n'étant possible que si, et seulement si, les deux expériences sont indépendantes.

Propriété 9

Soit $A_1, \dots, A_n$ *n* expériences quelconques. Nous avons l'inégalité

$$H(A_1, \dots, A_n) \leq \sum_{k=1}^{n} H(A_k)$$

L'égalité n'étant possible que si, et seulement si, les *n* expériences sont indépendantes.

Propriété 10

Soit A et B deux expériences quelconques. Nous avons l'égalité

$$H_B(A) = H_A(B) + H(A) - H(B)$$

Théorème d'unicité

Théorème 1

Soit $H(p_1, \dots, p_n)$ une fonction symétrique définie pour tout naturel *n* et pour toute répartition $p_1, \dots, p_n$. Supposons que pour tout *n* cette fonction soit continue par rapport à l'ensemble de ces variables et qu'elle satisfasse aux conditions suivantes :

1°) la fonction $H(p_1, \dots, p_n)$ est maximale pour la répartition uniforme $p_k = \frac{1}{n}, 1 \leq k \leq n$;

2°) si $p_1, \dots, p_n$ et $q_{k1}, \dots, q_{km}$, $1 \leq k \leq n$, sont n+1 répartitions et si nous construisons la nouvelle répartition $\pi_{kl} = p_k q_{kl}$, $1 \leq k \leq n$, $1 \leq l \leq m$, alors nous avons

$$H(\pi_{1l}, \dots, \pi_{nm}) = H(p_1, \dots, p_n) + \sum_{k=1}^{n} p_k H(q_{k1}, \dots, q_{km})$$

3°) $H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$

Dans ces conditions nous avons

$$H(p_1, \dots, p_n) = -\lambda \sum_{k=1}^{n} p_k \, log \, p_k$$

Où λ est une constante positive arbitraire.

Construction de la probabilité à l'aide de l'information

Soit $\Omega = \{\omega_1, \dots, \omega_n\}$ et $\wp(\Omega)$ l'espace probabilisable fini associé. Notons par $I(A)$ l'information fournie par l'événement $A \in \wp(\Omega)$.

Axiome 1 : (des valeurs extrêmes)

L'événement certain ne contient aucune information et l'événement impossible contient une information infinie :

$$I(\Omega) = 0, I(\emptyset) = +\infty$$

Axiome 2 : (de monotonie)

Si $A, B \in \wp(\Omega)$ et $B \subset A, B \neq A$, alors

$$I(A) < I(B)$$

Axiome 3 : (de la réunion)

Soit *f* une fonction convexe décroissante. Si $A, B \in \wp(\Omega)$ et $A \cap B = \emptyset$, alors

$$I(A \cup B) = f\left(f^{-1}\big(I(A)\big) + f^{-1}\big(I(B)\big)\right)$$

Théorème 2

Soit $\wp(\Omega)$ un espace probabilisable, I une fonction satisfaisant aux axiomes 1, 2 et 3, et supposons que la fonction f de l'axiome 3 soit continue et qu'elle vérifie les conditions

a) La fonction inverse f^{-1} est définie sur l'ensemble R_+ des réels non négatifs et prend ses valeurs dans $[0,1]$;
b) $f^{-1}(0) = 1, f^{-1}(+\infty) = 0$

Alors la fonction réelle *p* définie pour tout événement $A \in \wp(\Omega)$ par l'égalité

$$p(A) = f^{-1}\big(I(A)\big)$$

Est une probabilité définie sur l'espace probabilisable $\wp(\Omega)$.

Définition 5

Les événements $A, B \in \wp(\Omega)$ sont dits *informationnellement indépendants,* si on a

$$I(A \cap B) = I(A) + I(B)$$

Définition 6

Les événements $A_1, \dots, A_n \in \wp(\Omega)$ sont dits *informationnellement indépendants,* si pour tout $1 \leq i_1, < \cdots < i_l \leq n$ on a

$$I(A_{i1} \cap \dots \cap A_{il}) = I(A_{i1}) + \dots + I(A_{il})$$

Théorème 3

Soit $\wp(\Omega)$ un espace probabilisable, I une fonction satisfaisant aux axiomes 1, 2 et 3, et supposons que la fonction *f* de l'axiome 3 soit continue et qu'elle vérifie les conditions (a) et (b) du théorème 2. Soit ensuite *p* la probabilité qui correspond à I, et dont l'existence est assurée par le théorème 2.

Si l'indépendance informationnelle des événements de $\wp(\Omega)$ coïncide avec l'indépendance usuelle par rapport à la probabilité *p*, alors

$$f(x) = c\ log\frac{1}{x}, \qquad c > 0$$

Ce qui entraîne

$$p(A) = e^{-\frac{1}{c}I(A)}, \quad c > 0$$

ou

$$I(A) = -c \log p(A), \quad c > 0$$

Théorème 4

La répartition $p_1, \dots, p_n$ qui rend maximale l'information

$$H = -\sum_{k=1}^{n} p_k \, log \, p_k$$

En connaissant les moments

$$E\left(g_j \circ f_i\right) = \sum_{k=1}^{n} p_k \, g_l\left(x_{jk}\right), \qquad 1 \le j \le r, \quad 1 \le l \le m,$$

des variables aléatoires f_j, $1 \le j \le r$, où f_j prend les valeurs x_{jk}, $1 \le k \le n$,respectivement avec les probabilités p_k, $1 \le k \le n$, est donnée par

$$p_k = \frac{1}{\Phi(\lambda_{11}, \dots, \lambda_{mr})} e^{-\sum_{l=1}^{m} \sum_{j=1}^{r} \lambda_{lj} \, g_l\left(x_{jk}\right)}, \quad 1 \le k \le n$$

avec

$$\Phi(\lambda_{11}, \dots, \lambda_{mr}) = \sum_{k=1}^{n} e^{-\sum_{l=1}^{m} \sum_{j=1}^{r} \lambda_{lj} \, g_l\left(x_{jk}\right)}$$

Les grandeurs $\lambda_{11}, \dots, \lambda_{mr}$ résultent ensuite des égalités

$$E\left(g_l \circ f_j\right) = -\frac{\partial \log \Phi(\lambda_{11}, \dots, \lambda_{mr})}{\partial \, \lambda_{lj}}, \qquad 1 \le j \le r, \quad 1 \le l \le m.$$

Transmission de l'information sans codage

Il est admis que tout système de communication, de transmission de l'information, quelle que soit sa nature, rentre dans le schéma général présenté dans la figure ci-après.

L'information se transmet à l'aide de signaux qui sont émis par une source.

Ces signaux peuvent être transformés par codage en d'autres signaux.

Les signaux se propagent par une voie ou canal de transmission, où ils peuvent subir différentes altérations.

Ils arrivent à l'utilisateur, où un traducteur assure le décodage permettant d'obtenir les signaux émis au début.

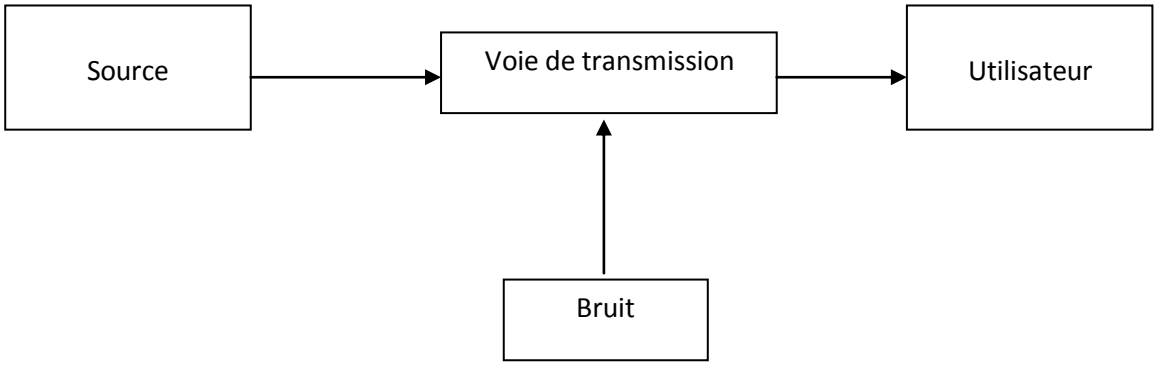


Figure 1 : Schéma d'un système de transmission de l'information sans codage

Nous nous posons le problème d'associer un modèle mathématique au schéma général de transmission de l'information :

Soit X l'ensemble fini des signaux qui peuvent être émis. Soit Y l'ensemble des signaux reçus.

Définition 7

L'espace probabilisé $\{X, x, p(x)\}$ s'appelle *source*.

Définition 8

Le triplet $[X, p(y/x), Y]$ formé des deux ensembles X et Y et de la probabilité conditionnelle $p(y/x)$ définie pour tout $x \in X$ et $y \in Y$ s'appelle *voie* ou *canal de transmission*. Si $p(y/x)$ prend seulement les valeurs 0 et 1 pour tout $x \in X$ et $y \in Y$, la voie est dite *sans bruit* ; sinon, elle est dite *avec bruit.*

Définition 9

L'espace probabilisé $\{Y, y, p(y)\}$ s'appelle *utilisateur*, la probabilité $p(y)$ étant calculée d'après la relation

$$p(y) = \sum_{x \in X} p(x)\, p(y/x)$$

Définition 10

Le triplet $[\{X, x, p(x)\}, [X, p(y/x), Y],\ \{Y, y, p(y)\}]$formé par une source, une voie et un utilisateur s'appelle *système de transmission de l'information sans codage*.

Notations :

Notons par $H(X/y)$ la quantité d'information qui doit être émise à la source pour recevoir un seul signal *y*.

$$H(X/y) = \sum_{x \in X} p(x/y) \log p(x/y)$$

Notons par $H(X/Y)$ la quantité moyenne d'information nécessaire pour recevoir l'ensemble des signaux Y.

$$H(X/Y) = \sum_{y \in Y} p(y)\ H(X/y)$$

Définition 11

On appelle *capacité* d'une voie de transmission la valeur

$$C = \max_{p(x)}(H(X) - H(X/Y))$$

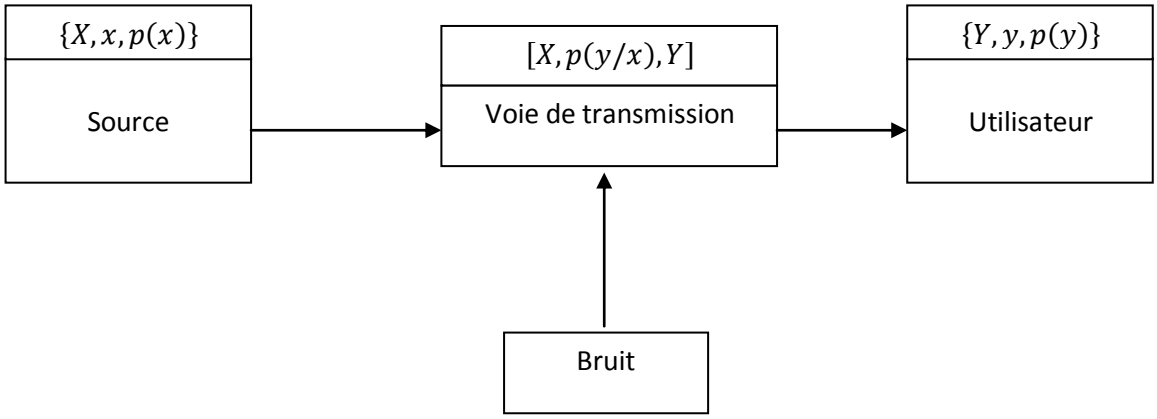


Figure 2 : Modèle mathématique d'un système de transmission de l'information sans codage

Codage

Notations

Notons par U l'ensemble de toutes les séquences de longueur *n* formées avec les signaux $x \in X$, c'est-à-dire

$$U = \prod_{j=1}^{n} X_j,\ \ X_j = X,\qquad 1 \le j \le n$$

Notons par V l'ensemble de toutes les séquences de longueur *n* formées avec les signaux $y \in Y$, c'est-à-dire

$$V = \prod_{j=1}^{n} Y_j,\ \ Y_j = Y,\qquad 1 \le j \le n$$

Définition 12

L'espace probabilisé $\{U, u, p(u)\}$ s'appelle *la $n^{ième}$ extension de la source* $\{X, x, p(x)\}$.

Définition 13

Le triplet $[U, p(v/u), V]$ formé des deux ensembles $U = \prod_{j=1}^{n} X_j$, $X_j = X, 1 \leq j \leq n$ et $V = \prod_{j=1}^{n} Y_j$, $Y_j = Y, 1 \leq j \leq n$ et de la probabilité conditionnelle $p(v/u)$ définit pour tout $u \in U$ et $v \in V$ par $p(v/u) = p(y_1/x_1) \dots p(y_n/x_n)$, s'appelle *la n^ième^ extension de la voie de transmission* $[X, p(y/x), Y]$.

Définition 14

L'espace probabilisé $\{V, v, p(v)\}$ s'appelle *la n^ième^ extension de l'utilisateur* $\{Y, y, p(y)\}$, la probabilité $p(v)$ étant calculée d'après la relation

$$p(v) = \sum_{u \in U} p(u)\, p(v/u)$$

Définition 15

Le triplet $[\{U, u, p(u)\}, [U, p(v/u), V], \{V, v, p(v)\}]$ s'appelle *la n^ième^ extension du système de transmission* $[\{X, x, p(x)\}, [X, p(y/x), Y], \{Y, y, p(y)\}]$

Théorème 5

Soit $\{X, x, p(x)\}$ une voie de transmission de capacité C. alors sa *n^ième^ extension* $[U, p(v/u), V]$ a la capacité *nC.*

Théorèmes de codage sans bruit

Définition 16

On appelle code instantané un code dans lequel les différentes séquences de codage peuvent avoir des longueurs différentes et dans lequel aucune séquence de codage n'est le '*préfixe*' d'une autre séquence de codage.

Lemme 1

Pour que le codage instantané soit possible à l'aide d'une famille de N séquences de signaux simples, de longueurs $n_1, \dots, n_N$, que l'on fait correspondre aux signaux initiaux $x_1, \dots, x_N$, il est nécessaire et suffisant que l'inégalité suivante soit vérifiée

$$\sum_{j=1}^{N} Q^{-nj} \leq 1$$

où Q est le nombre total des signaux simples.

Lemme 2

Soient $c_1, \dots, c_n$ des nombres quelconques et $q_1, \dots, q_n$ des nombres négatifs tels que

$$\sum_{i=1}^{n} q_i = 1$$

Lemme 3

Soient $p_1, \dots, p_n$ et $q_1, \dots, q_n$ deux répartitions. Alors

$$-\sum_{i=1}^{n} q_i \, log \, q_i \leq -\sum_{i=1}^{n} q_i \, log \, p_i$$

Théorème 6

Pour pouvoir effectuer un codage instantané en utilisant Q signaux simples, à l'aide desquels il faut transmettre une quantité d'information $H(X)$, il est nécessaire que la longueur moyenne des séquences de codage, attachées aux signaux initiaux de l'ensemble X, ne soit pas inférieure au nombre

$$\frac{H(X)}{\log Q}$$

Théorème 7

Soit X un ensemble de signaux initiaux à l'aide desquels nous transmettons une quantité d'information $H(X)$ et supposons que l'on effectue le codage instantané de cette information, à l'aide de l'ensemble A formé de Q signaux simples, en attachant des séquences de codage non pas à chaque signal initial mais directement aux blocs de M signaux initiaux. Alors la longueur moyenne des séquences de codage, divisée par M, s'approche de la borne inférieure $\frac{H(X)}{\log Q}$ lorsque M est suffisamment grand.

Codage avec bruit

Définition 17

Soit *e* un réel $0 < e < \frac{1}{2}$. Un ensemble de signaux $x_1, \dots, x_N$ de X constitue un *ensemble maximal*, noté $M[e]$, si

1°) à chaque signal x_i, $1 \leq i \leq N$, correspond un ensemble de signaux *y*, noté B_i, tel que

$$p(B_i/x_i) \geq 1 - e$$

2°) les ensembles $B_1, \dots, B_N$ sont deux à deux disjoints ;

3°) il n'est pas possible de trouver un signal x_{N+1} et un ensemble B_{N+1} tel que système $x_1, \dots, x_N, x_{N+1}$ et la famille correspondante d'ensembles $B_1, \dots, B_N, B_{N+1}$ vérifient les conditions 1 et 2.

Définition 18

Un ensemble maximal $M[e]$ est *L-borné,* $0 < L < 1$*,* par rapport aux probabilités d'émission $p(x)$, si les ensembles B_i, $1 \leq i \leq N$, vérifient la relation

$$p(B_i) < L, \;\; 1 \leq i \leq L$$

la probabilité de l'ensemble B_i étant égale à

$$p(B_i) = \sum_{x \in X} p(x, B_i) = \sum_{x \in X} p(x) p(B_i/x), \;\; 1 \leq i \leq N$$

Définition 19

Un ensemble de signaux $x_1, \dots, x_N$ de l'ensemble X, muni d'une probabilité $p(x)$, constitue un ensemble *étendu* par rapport aux réels *e* et *K*, $0 < e < 1, 0 < K < 1$, et sera noté $M[e, K]$ si

1°) à chaque signal x_i, $1 \leq i \leq N$, correspond un ensemble A_i de signaux de Y, tel que

$$p(A_i/x_i) \geq 1 - e, \;\; 1 \leq i \leq N$$

2°) les probabilités des ensembles $A_1, \dots, A_N$ vérifient l'inégalité

$$p(A_i) < K, \;\; 1 \leq i \leq N$$

où

$$p(A_i) = \sum_{x \in X} p(x, A_i) = \sum_{x \in X} p(x) p(A_i/x)$$

Lemme 4

Supposons que l'ensemble X de signaux initiaux contient un ensemble maximal K-borné $M[e]$,

$$M[e] = \{x_1, \dots, x_N\}$$

Ainsi qu'un ensemble étendu $M[e, K]$,

$$M[e, K] = \{x_1', \dots, x_M'\}$$

où $0 < a < 1, \;\; 0 < e < 1, \;\; 0 < K < 1$, et

$$M[e] \cap M[e, K] = \emptyset$$

alors

$$NK > (e - a) p(M[e, K]) + (1 - e) p(M[e])$$

Lemme 5

Pour tout $\varepsilon > 0$ et $\delta > 0$, il existe un naturel $n_0(\varepsilon\,, \delta)$tel que $n \geq n_0(\varepsilon\,, \delta)$ implique

$$p\left(\left|\frac{1}{n}\log p(u) + H(X)\right| \geq \varepsilon\right) \leq \delta$$

et il existe un naturel $n_0^*(\varepsilon\,, \delta)$ tel que $n \geq n_0^*(\varepsilon\,, \delta)$ implique

$$p\left(\left|\frac{1}{n}\log p(u/v) + H(X/Y)\right| \geq \varepsilon\right) \leq \delta$$

Lemme 6

Soit $Z \subset U \times V$ l'ensemble des couples de signaux à l'entrée et à la sortie de la n^{ième} extension d'un système de transmission de l'information défini par

$$Z = \{(u, v) : p(u, v) > 1 - \delta_2\}$$

Soit ensuite $U_0 \subset U$ un ensemble de signaux à l'entrée ; tel que

$$p(U_0) > 1 - \delta_2$$

Pour chaque $u \in U$, soit A_u l'ensemble des éléments *v* tels que $(u, v) \in Z$.

Soit, enfin, $U_1 \subset U_0$ le sous ensemble des signaux *u* qui vérifient la condition

$$p(A_n/u) \geq 1 - a$$

Alors on a

$$p(U_1) > 1 - \delta_2 - \frac{\delta_1}{a}$$

Théorème 8

Soit une voie de transmission $[X, p(y/x), Y]$ de capacité C. Soit ensuite H et *e* donnés, tels que $0 < H < C$ et $e > 0$.

Il existe alors un naturel $n_0(e, H)$, qui dépend de *e* et de H ; tel que dans chaque n^{ième} extension de cette voie $[U, p(v/u), V]$ avec $n > n_0(e, H)$, il existe un ensemble de signaux $u_1, \ldots, u_N$, avec $N > 2^{nH}$, à chaque signal u_k étant associé un ensemble A_k d'éléments *v*, $1 \leq k \leq N$, ces ensembles étant deux à deux disjoints et tels que

$$p(A_i/u_i) \geq 1 - e, \quad 1 \leq i \leq N$$

Divergence de Kullback

Considérons les espaces probabilisés

$$(\mathcal{X}, \mathfrak{S}, \mu_i), \qquad i = 1,2$$

Où $\mathcal{X}$ est un ensemble d'éléments, $\mathfrak{S}$ est la σ −algèbre de tous les événements possibles construits avec les éléments de $\mathcal{X}$, qui est muni de mesures de probabilité μ_i, $i = 1,2$.

Les éléments de $\mathcal{X}$ peuvent être univariés ou multivariés, discrets ou continus, qualitatifs ou quantitatifs.

Nous supposerons que les mesures de probabilité μ_1 et μ_2 sont absolument continues l'une par rapport à l'autre. Nous noterons : $\mu_1 \equiv \mu_2$.

Soit λ une mesure de probabilité telle que $\lambda \equiv \mu_1$ et $\lambda \equiv \mu_2$.

D'après le théorème de Radon-Nikodym, il existe des fonctions $f_i(x), i = 1,2$, appelées densités de probabilité généralisées, λ-mesurables, unique presque partout, $0 < f_i(x) < \infty \ [\lambda], i = 1,2$ telles que :

$$\mu_i(E) = \int_E f_i(x) d\lambda(x), \qquad i = 1,2$$

Pour tout $E \in \mathfrak{S}$.

Notons par X la variable aléatoire et par x la valeur spécifique que peut prendre la variable aléatoire X.

Définition 1

Soit $H_i, \ i = 1,2$ l'hypothèse que X provient de la population statistique dont la mesure de probabilité est μ_i. On appelle information en $X = x$ pour la discrimination en faveur de H_1 contre H_2 la quantité :

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1/x)}{P(H_2/x)} - \log \frac{P(H_1)}{P(H_2)} \ [\lambda]$$

Où $P(H_i), \ i = 1,2$ sont les probabilités a priori de H_i et $P(H_i/x)$ sont les probabilités a posteriori de H_i (ou probabilités conditionnelles de H_i sachant $X = x$).

Définition 2

On appelle information moyenne par observation, par rapport à μ_1, pour la discrimination en faveur de H_1 contre H_2 la quantité :

$$I(1:2;E) = \frac{1}{\mu_1(E)} \int_E \log \frac{f_1(x)}{f_2(x)} d\mu_1(x)$$

$$= \frac{1}{\mu_1(E)} \int_E f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x), \quad \mu_1(E) > 0$$

$$= 0, \qquad \mu_1(E) > 0$$

Avec $d\mu_1(x) = f_1(x)\, d\lambda(x)$.

Lorsque E est l'espace $\mathcal{X}$ en entier, l' information moyenne par observation, par rapport à μ_1, pour la discrimination en faveur de H_1 contre H_2 s'écrit :

$$I(1:2) = \int \log \frac{f_1(x)}{f_2(x)} d\mu_1(x) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x)$$

$$= \int \log \frac{P(H_1/x)}{P(H_2/x)}\, d\mu_1(x) - \log \frac{P(H_1)}{P(H_2)}$$

Définition 3

- La quantité :

$$I(2:1) = \int f_2(x) \log \frac{f_2(x)}{f_1(x)} d\lambda(x)$$

Définit l'information moyenne par observation, par rapport à μ_2, pour la discrimination en faveur de H_2 contre H_1

- La quantité :

$$-I(2:1) = \int f_2(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x)$$

Définie l'information moyenne par observation, par rapport à μ_2, pour la discrimination en faveur de H_1 contre H_2.

Définition 4

La divergence entre les hypothèses H_1 et H_2, ou entre les mesures μ_1 et μ_2, est définie par :

$$J(1,2) = I(1:2) + I(2:1)$$

$$= \int \big(f_1(x) - f_2(x)\big) \log \frac{f_1(x)}{f_2(x)} d\lambda(x)$$

$$= \int \log \frac{P(H_1/x)}{P(H_2/x)}\, d\mu_1(x) - \int \log \frac{P(H_1/x)}{P(H_2/x)}\, d\mu_2(x)$$

CHAPITRE III

APPLICATION DE LA THÉORIE MATHÉMATIQUE DE L'INFORMATION POUR L'ÉLABORATION DE QUESTIONNAIRES

Résumé

Dans ce chapitre nous étudions le moyen de mesurer quantitativement l'information apportée par l'observation lors de l'identification des lois qui régissent les phénomènes aléatoires.

A partir de l'étude sur une variable, nous construisons le concept de gain d'information sur la notion d'entropie relative.

Nous démontrons que, lors de l'affinement de l'étude par désagrégation des modalités du caractère, il existe un seuil pour les probabilités rattachées aux différentes modalités. Ce seuil détermine les situations où le gain d'information est définitif et celles où il est illusoire.

Nous montrons comment cette étude peut s'étendre au cas de plusieurs variables. Nous en déduisons une méthode quantitative de sélection, pas à pas, de variables, respectant le principe de l'entropie maximale. Cette méthode aboutit à l'élaboration, après pré enquête, de questionnaires parcimonieux susceptibles de récolter la plus grande part d'information.

1. Introduction

Tel que relatée par Segal [37], mesurer quantitativement l'information apportée par un échantillon est une préoccupation qui a commencé depuis les travaux de Fisher [1, 5, 7, 11].

Sa définition de l'information s'inscrit au sein d'une structure statistique paramétrée et concerne avant tout ce que l'échantillon peut enseigner à propos de la valeur du paramètre.

En 1948, initiée par Shannon [38, 38bis], une théorie mathématique de l'information, en tant que notion probabiliste, a vu le jour. Cette théorie a suscité beaucoup d'attention et a été l'objet de beaucoup de recherches dont ceux de Khintchin [17] et de Kolmogorov [19, 19bis].

Les liens entre les différentes définitions de l'information ont fait l'objet de recherches inaugurant une théorie de l'information unifiée, à la fois statistique et probabiliste (Schützenberger [21bis, 22, 22bis, 22ter], Kullback [25]).

Kullback et Leibler [24], dans leur travail de réinterprétation des statistiques autour de la notion d'information, produisent la définition de l'information moyenne apportée par un échantillon en

faveur d'une hypothèse H_1 contre une hypothèse H_2 et la définition de la divergence, ou entropie relative, entre deux distributions de probabilités.

Dans notre travail, nous nous plaçons dans le cas où nous ne possédons aucune information préalable et, partant du principe de la raison insuffisante, nous admettons que la première distribution de probabilités est la distribution uniforme. La deuxième distribution est celle qui régie le phénomène aléatoire et que nous cherchons à identifier. De cette manière, l'entropie relative s'identifie, à la limite, avec l'information apportée par l'échantillon.

Nous étudions alors le moyen de concevoir les différentes modalités des variables afin de respecter le principe du maximum d'entropie (Jaynes[26 , 26bis]).

Comme résultat pratique, nous aboutissons à une méthode de classification descendante à croissance d'information quasi-monotone et à une méthode de conception de questionnaires évitant le plus possible l'acquisition d'information illusoire.

2. Entropie de la distribution a priori (Capacité)

Un caractère se répartit en modalités qui peuvent être en nombre fini ou infini.

Le caractère peut être une variable aléatoire continue.

Dans certains cas, c'est l'observation qui permet d'identifier les différentes modalités du caractère. Dans d'autres cas, ils peuvent déjà être déterminés avant l'observation.

Le rôle de l'observation est de noter les occurrences de chaque modalité.

L'incertitude sur l'issue de l'expérience est fonction des probabilités rattachées aux différentes modalités (dont le nombre est k). Elle est mesurée par la quantité

$$H(X) = -\sum_{i=1}^{k} p_i Log\ p_i$$

qui est appelée entropie de l'expérience.

La quantité d'information récoltée à la suite de l'expérience est définie comme étant la part d'incertitude éliminée.

Propriété

L'entropie $H(X)$ est la quantité totale d'information susceptible d'être récoltée. Elle atteint son maximum lorsque toutes les probabilités $p_i \quad i = 1,\ldots, k$ sont égales.

Elle vaut alors

$$H(X) = Log\, k\ .$$

Démonstration

D'une part,

$$H\left(\frac{1}{k},\ldots,\frac{1}{k}\right) = -\sum_{i=1}^{k} \frac{1}{k}.Log\frac{1}{k} = -\frac{k}{k}.Log\frac{1}{k} = Logk$$

D'autre part, nous allons utiliser l'inégalité de Jensen qui s'énonce de la manière suivante :

Si f est une fonction concave sur l'intervalle $[a\,,b]$, et $x_1, \dots, x_k$ sont k valeurs arbitraires de l'argument x, alors pour tous nombres positifs $\lambda_1, \dots, \lambda_k$ dont la somme est égale à 1, on a l'inégalité

$$\sum_{i=1}^{k} \lambda_i . f(x_i) \leq f\left(\sum_{i=1}^{k} \lambda_i . x_i\right)$$

appliquons cette inégalité de Jensen pour

$$x_i = p_i\,, \qquad \lambda_i = \frac{1}{k}\ , \qquad 1 \leq i \leq k,$$

$$f(x) = -xLogx$$

nous obtenons

$$-\sum_{i=1}^{k} \frac{1}{k} . p_i . \mathrm{Log}\, p_i \leq -\left(\sum_{i=1}^{k} \frac{1}{k} . p_i\right) Log\left(\sum_{i=1}^{k} \frac{1}{k} . p_i\right)$$

et puisque

$$\sum_{i=1}^{k} p_i = 1$$

il résulte

$$\frac{1}{k} H(p_1, \dots, p_k) \leq -\frac{1}{k}\left(\sum_{i=1}^{k} p_i\right) Log\left(\frac{1}{k} . \sum_{i=1}^{k} p_i\right) = -\frac{1}{k} . Log\frac{1}{k}$$

c'est-à-dire

$$H(p_1, \dots, p_k) \leq -Log\frac{1}{k} = \mathrm{Log}\, k \qquad \blacksquare$$

Ainsi, quelle que soit sa distribution de probabilités, une variable aléatoire à k modalités ne peut pas produire une quantité d'information supérieure à $Log\, k$.

$Log\, k$ peut être vu comme la capacité d'une variable aléatoire à k modalités.

Plus le nombre de modalités est grand et plus cette capacité est grande. Augmenter le nombre de modalités prépare à recevoir une plus grande quantité d'information.

Etant logarithmique, la croissance de la capacité est rapide au départ mais elle devient insignifiante au-delà d'un certain rang. A la limite, elle converge vers zéro.

Si, avant l'observation, nous ne connaissons du caractère que le nombre k de modalités, sans autre information préalable, la distribution des probabilités est uniforme.

$Log\, k$ peut ainsi être également vu comme étant l'entropie rattachée à cette distribution a priori.

3. Production de l'expérience – Entropie de la distribution a posteriori

Lorsque l'expérience est entreprise, à chaque modalité M_i se trouvera associée une fréquence relative f_i.

Pendant que le nombre d'observation augmente, les fréquences relatives évoluent. Elles se stabilisent graduellement et finissent, en vertu de la loi des grands nombres, par converger vers des constantes $p_i,\ i = 1, \dots, k$ qui sont les probabilités.

L'entropie de l'expérience, dans les étapes intermédiaires, se mesure par :

$$H_f(X) = -\sum_{i=1}^{k} f_i \log f_i$$

Cette quantité, évoluant suivant l'évolution des $f_i,\ i = 1, \dots, k$ converge vers la quantité

$$H(X) = -\sum_{i=1}^{k} p_i \log p_i$$

H(X) est l'entropie rattachée à la distribution a posteriori $p_i,\ i = 1, \dots, k$.

4. Notion de gain d'information (Entropie Relative).

L'entropie a posteriori H(X) est nécessairement inférieure à l'entropie a priori :

$$H(X) \leq \log k$$

L'incertitude a priori, qui est l'incertitude totale, ne peut pas être entièrement éliminée par l'expérience. Elle est la somme de deux incertitudes dont une seulement peut disparaître grâce à l'expérience.

La partie incompressible est l'entropie H(X) spécifique à la distribution que nous cherchons à identifier. L'autre partie est l'écart entre H(X) et $\log k$ et c'est la quantité possible d'information qui peut être apportée par l'expérience. Nous l'appellerons gain de l'information, GI :

$$GI = \log k - H(X)$$

5. Convergence du GI vers une constante lorsque le nombre d'observations tend vers l'infini

Pour un nombre d'observations N donné, le Gain d'Information est :

$$GI_f = \log k - H_f(X)$$

Avec l'augmentation du nombre d'observations, la quantité GI_f converge vers la quantité GI.

$$\lim_{N\to\infty} GI_f = \lim_{N\to\infty} \left[\log k - H_f(X)\right] = \log k - H(X) = GI$$

GI est en fait la distance de Kullback – Leibler, ou entropie relative, entre la distribution uniforme et la distribution à identifier.

6. Convergence du GI vers une constante lorsque le nombre des modalités tend vers l'infini

Le GI converge également vers une constante quand le nombre de modalités k tend vers l'infini. Ceci peut être perçu comme une version du théorème central limite.

Pour fixer les idées, considérons la loi binomiale B(k,p), k étant le nombre de modalités et p la probabilité de succès dans l'épreuve de Bernoulli rattachée à cette loi binomiale.

Pour chaque k et p fixés, la variable binomiale B(k,p) se trouve associée à un GI qui vaut :

$$GI_k = \log k - H_k(X) = \log k + \sum_{i=1}^{k} C_n^i p^i (1-p)^{n-i} \log\left[C_n^i p^i (1-p)^{n-i}\right]$$

Lorsque k augmente, la convergence de la loi binomiale vers la loi normale entraîne la convergence du GI vers une constante qui est le gain d'information associé à la loi normale.

D'une manière générale, il est évident que lorsqu'une loi converge vers une autre loi son GI converge vers le GI de cette loi.

7. Désagrégation de modalités

7.1. Augmentation de l'entropie

Théorème 1

Si X est un caractère à k modalités ayant respectivement les probabilités $p_1, p_2, \ldots, p_k$ et si nous construisons un nouveau caractère X' en désagrégeant l'une quelconque des modalités, M_{i_0}, ayant la probabilité p_{i_0}, en deux modalités M_{i_1} et M_{i_2},

avec pour probabilités respectives p_{i_1} et p_{i_2} telles que $p_{i_0} = p_{i_1} + p_{i_2}$ alors

$$H(X') \geq H(X)$$

Autrement dit, l'entropie augmente toujours en désagrégeant les modalités.

7.2. Cas différents de l'augmentation ou bien de la diminution du GI :

Théorème 2

Il existe une condition pour que le GI augmente en désagrégeant une modalité M_{i_0} (de probabilité p_{i_0}) en deux modalités M_{i_1} et M_{i_2} de probabilités respectives $c.p_{i_0}$ et $(1-c)p_{i_0}$, où c est un réel arbitraire compris entre 0 et 1.

Cette condition est :

$$p_{i_0} \leq \frac{Log(k+1) - Logk}{c.Log\frac{1}{c} + (1-c).Log\frac{1}{1-c}}$$

Démonstration

Nous avons

$$GI_{k+1} = Log\,(k+1) - H(X')$$

et

$$GI_k = Log\,k - H(X)$$

Par soustraction

$$GI_{k+1} - GI_k = [Log\,(k+1) - Log\,k] - [H(X') - H(X)]$$

$$= Log\frac{k+1}{k} - p_{i_0}\left[c.Log\frac{1}{c} + (1-c)Log\frac{1}{1-c}\right]$$

La différence $GI_{k+1} - GI_k$ n'est positive que si

$$p_{i_0} \leq \frac{Log\,(k+1) - Log\,k}{c.Log\frac{1}{c} + (1-c).Log\frac{1}{1-c}} = \alpha$$ ■

Discussion

Lorsqu'il s'agit d'augmenter le nombre de modalités d'un caractère, nous pouvons rencontrer, à chaque désagrégation, deux situations possibles :

- La probabilité p_{i_0} rattachée à cette modalité est inférieure à α . Dans ce cas nous obtenons une croissance du GI.
- La probabilité p_{i_0} est supérieure à α , dans quel cas le GI diminuera.

La décomposition dans le second cas apportera une information négative en augmentant l'incertitude globale.

En multipliant les décompositions des modalités entrant dans le second cas, nous ferons augmenter le GI rapidement et d'une manière illusoire avec l'augmentation de k. Il diminuera par la suite lorsque nous devons désagréger les modalités ayant des probabilités entrant dans le premier cas.

Pour éviter cette situation et, disposant d'information partielle, il faut choisir comme modalités à désagréger celles dont les probabilités conduisent à maximiser l'incertitude concernant l'information manquante.

Le processus doit être conduit de sorte à désagréger d'abord les modalités dont les probabilités p_{i_0} vérifient la condition $p_{i_0} \geq \alpha$.

Lorsqu'à une étape, il n'y a pas de probabilités vérifiant cette condition, alors l'augmentation du GI, suite à la décomposition, est réellement définitive.

Remarque : Distance de Kullback – Leibler comme mesure d'hétérogénéité

Le gain d'information GI, qui peut être également écrit sous la forme

$$GI = Log\ k.\prod_{i=1}^{k} p_i^{\,p_i}$$

peut servir comme mesure d'hétérogénéité de la distribution de probabilités.

Cette quantité est maximale lorsque l'une des probabilités est égale à 1, et elle est minimale lorsque toutes les probabilités sont égales. Elle peut servir comme mesure de dispersion lorsqu'il s'agit d'un caractère qualitatif ordinal.

8. Cas de plusieurs variables

Théorème 3

La capacité d'un groupe de N variables binaires est égale à N.

Plus généralement, la capacité d'un groupe de N variables, ayant respectivement $m_1, m_2, \ldots, m_N$ modalités, est égale à : $\sum_{i=1}^{N} Log\ m_i$

Démonstration

Il faut souligner d'abord que la description d'une population à l'aide de N variables peut se ramener à la description à l'aide d'une seule variable.

Si les N variables sont binaires, la nouvelle variable possédera 2^N modalités, sa capacité est

$$Log\ 2^N = N\ .$$

Si les N variables ont respectivement $m_1, m_2, \ldots, m_N$ modalités, la nouvelle variable possédera $m_1.m_2.\ldots.m_N$ modalités et sa capacité est

$$Log\ m_1.m_2.\ldots.m_N = \sum_{i=1}^{N} Log\ m_i$$

En effet, du fait qu'il y ait N variables, chaque individu est identifié par un vecteur à N composantes.

Si les variables sont binaires, il existe 2^N vecteurs différents. Chacun de ces vecteurs peut être considéré comme une modalité de la nouvelle variable.

Si les variables ont des nombres différents de modalités $m_1, m_2, \ldots, m_N$, il existe $m_1.m_2.\ldots.m_N$ vecteurs différents qui seront autant de modalités pour la nouvelle variable.

9. Classification descendante à augmentation monotone du GI

Lorsque nous nous disposons à répartir une population en classes, en disposant pour cela d'un certain nombre V de variables ayant chacune un nombre de modalités déterminé, si notre information préalable, résultant d'une pré enquête, est insuffisante, nous pouvons procéder à une sélection graduelle et parcimonieuse des variables en respectant la discussion précédente et le théorème ci-dessus.

- La première variable sélectionnée est celle dont la distribution des fréquences relatives est la plus uniforme possible.
- La deuxième variable, ajoutée à la première conduit à une nouvelle distribution de fréquences relatives (qui seront assimilées à des probabilités).
  Sur les (V-1) possibilités de choix, sélectionner celle qui produit la distribution de probabilités vérifiant la condition α ou, sinon, en être la plus proche.

- Le même procédé permettra de sélectionner une à une les variables suivantes.

Nous obtenons de cette manière un classement décroissant des V variables en fonction de leur pouvoir discriminant.

- Dans le cas où notre but est la subdivision de la population sous forme de classification descendante :
  Si nous désirons répartir cette population en M classes, alors nous nous limitons aux L premières variables (dont les nombres de modalités sont respectivement $m_1, m_2, \ldots, m_L$) et de sorte que $m_1.m_2.\ldots.m_L$ soit le proche de M (par excès).

  Par le fait que l'information acquise est la moins illusoire possible, une partition conçue de la sorte sera la plus stable.

  Dans le cas où nous désirons élaborer un questionnaire :

  Nous pouvons nous limiter au nombre de variables qui assure le pourcentage attendu de la quantité d'information à récolter.

  Si, ultérieurement, le nombre des observations devra augmenter, le déséquilibre entre les fréquences relatives ne peut que s'accentuer. C'est ce qui assurera un supplément de gain effectif d'information.

CHAPITRE IV

ELÉMENTS DE RÉSEAUX DE NEURONES

Introduction

Inspirés par la structure du cerveau humain, les réseaux de neurones artificiels ont été beaucoup appliquées dans des domaines tels que la reconnaissance des formes, l'optimisation, le codage, le contrôle … etc., principalement à cause de leur aptitude à résoudre des problèmes classés difficiles. Cette aptitude leurs est prodiguée par le fait qu'ils apprennent directement et automatiquement des données, sans besoin de formulation a priori.

Un réseau de neurone est généralement constitué d'un grand nombre de processeurs simples, les neurones, qui sont reliés les uns aux autres par des connections. Il apprend à résoudre des problèmes en ajustant les poids des interconnections d'une manière adéquate et en conformité avec les données. Plus que cela, le réseau de neurone, en apprenant, s'adapte facilement aux environnements nouveaux et peut traiter l'information bruitée, inconsistante, vague ou sujette à aléa.

L'hypothèse de base qui a présidé à l'émergence des réseaux de neurones, à l'image du cerveau, est que le comportement intelligent vient de la *structure* d'ensemble d'une part et, d'autre part du *comportement* des éléments de base. L'intérêt des neurones réside dans les propriétés qui résultent de leur association en réseaux. C'est l'architecture même du réseau, combinant le comportement des processeurs de base, qui lui octroie la capacité d'exécuter des fonctions jusqu'alors inédites telles que l'adaptation ou l'apprentissage.

Mais vus sous la perspective de la science statistique, les réseaux de neurones peuvent être perçus comme des extensions des techniques conventionnelles qui ont été développées depuis des décennies.

Aujourd'hui, avec la maturité grandissante du domaine, les qualités des neurones artificiels qui, au début s'apparentaient vaguement à celles des neurones biologiques, bénéficient de fondements statistiques solides.

C'est sous cet éclairage que nous allons d'abord décrire ce qu'est un neurone artificiel et ce que peut être son comportement. Nous étudierons ensuite différents types de réseaux, leurs architectures et leurs propriétés.

1. Différents neurones

Neurone biologique

Le neurone biologique est une cellule vivante spécialisée dans le traitement des signaux électriques, c'est une cellule composée d'un corps cellulaire et d'un noyau.

La structure d'un neurone se compose de trois parties :

- **La somma** : ou cellule d'activité nerveuse, au centre du neurone.
- **L'axone** : attaché au somma qui est électriquement actif, ce dernier conduit l'impulsion produite par le neurone.
- **Dendrites** : électriquement passives, elles reçoivent les impulsions d'autres neurones.

C'est par les dendrites que l'information pénètre de l'extérieur vers le soma, corps du neurone. L'information traitée par le neurone chemine ensuite le long de l'axone (unique) pour être transmise aux autres neurones.

La transmission entre deux neurones n'est pas directe. En fait, il existe un espace intercellulaire entre l'axone du neurone qui émet et les dendrites du neurone qui reçoit. La jonction entre deux neurones est appelée la synapse (fig. 1).

Les neurones font une sommation des signaux reçus en entrée et en fonction du résultat obtenu vont fournir un signal en sortie.

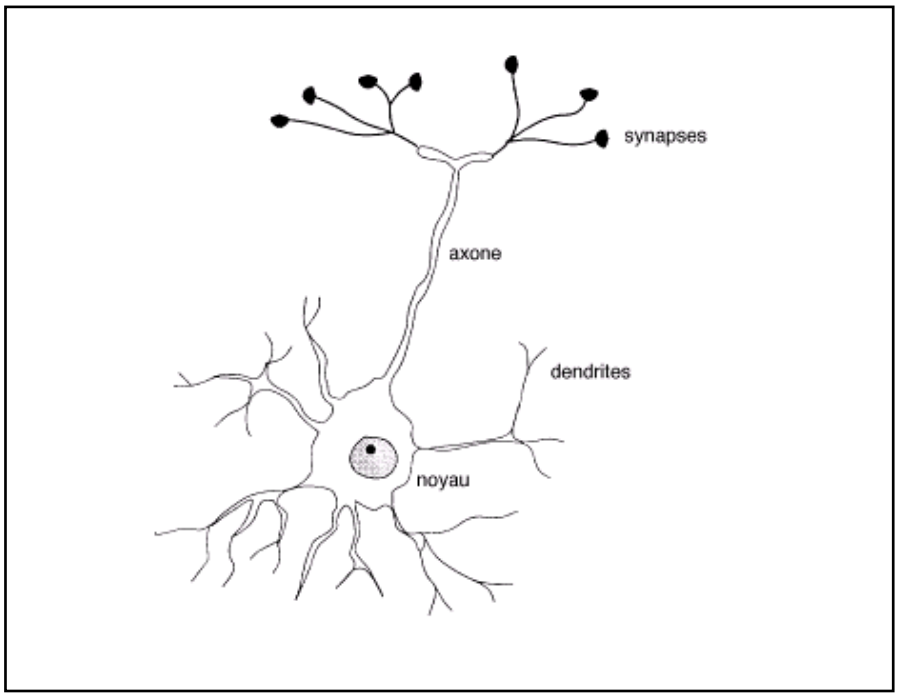


Figure 4.1 : le neurone biologique

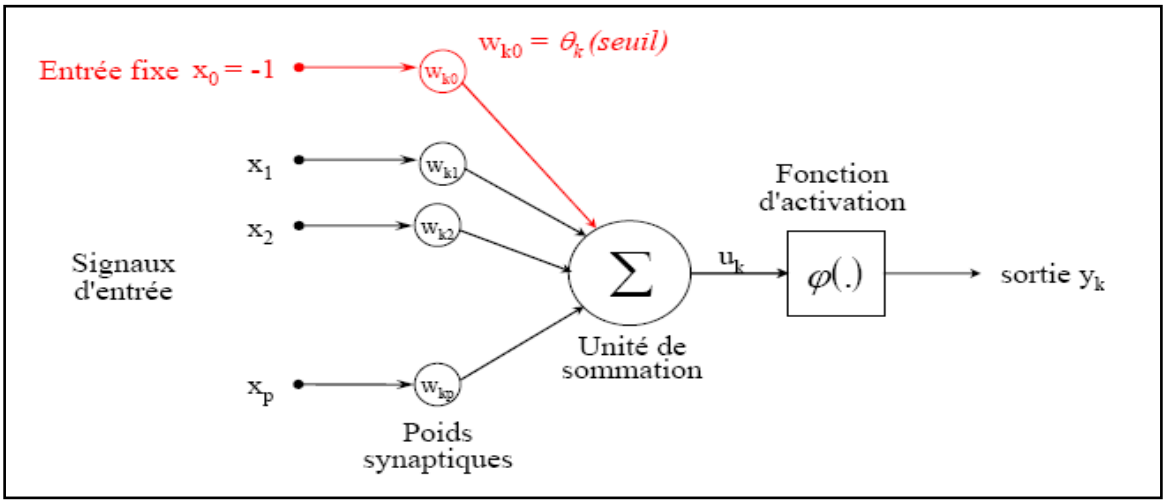
Neurone artificiel

La correspondance entre le neurone biologique et le neurone artificiel est modélisée dans le tableau 1 ci après

Neurone biologique	Neurone artificiel
Synapses	Poids de connexions
Axones	Signal de sortie
Dendrite	Signal d'entrée
Somma	Fonction d'activation

Tableau 1

Le fonctionnement d'un neurone artificiel peut être modélisé par le schéma suivant :



Ce modèle est mathématiquement décrit par deux équations :

$$u_k = \sum_{j=1}^{p} w_{kj}\, x_j \quad \text{et} \quad y_k = \varphi(u_k - \theta_k)$$

Où : $w_{k1}, w_{k2}, \ldots, w_{kp}$ sont les poids synaptiques du neurone *k,*

u_k est la sortie de l'unité de sommation,

θ_k est le seuil,

$\varphi(.)$ est la fonction d'activation,

y_k est le signal de sortie du neurone *k*.

Sur ce schéma, le neurone a plusieurs connexions en entrée le reliant à autant de neurones (ou entrées). Il reçoit de l'information provenant de chacun de ces neurones (ou entrées).

Les valeurs qu'il reçoit ainsi en entrée par chacune de ses connexions sont respectivement notées $X_1, X_2, \ldots, X_n$: ce sont les **entrées** du neurone. (-1 est le signal constant réservé au biais).

Toutes les connexions n'ont pas une importance égale pour le neurone. Certaines sont plus importantes que d'autres. Le **poids** *w* affecté à chaque connexion mesure cette importance relative : le poids est proportionnel à l'importance de la connexion. La somme pondérée, appelée activation est calculée :

$$a = w_1 X_1 + w_2 X_2 + \ \ldots + w_n X_n$$

Le neurone effectue ensuite une opération qui dépend de ***a***. Cela revient à dire qu'il applique une fonction f à la valeur ***a***. Cette fonction f est appelée **fonction d'activation.**

Le choix de cette fonction *f* se révèle être un élément constitutif important des réseaux de neurones.

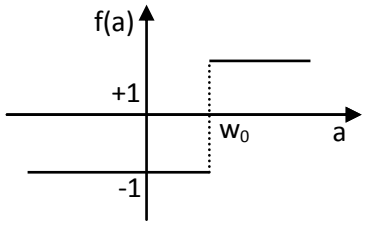
La valeur $f(a)$ calculée par le neurone constitue la **sortie** *S*. Cette valeur en sortie devient ensuite une valeur d'entrée pour tous les neurones avec lesquels notre neurone de référence est connecté.

2. Différentes fonctions d'activation

Il existe de nombreuses formes possibles pour la fonction d'activation. Les plus courantes sont celles qui suivent

a) Fonction seuil

$$f(a - w_0) = \begin{cases} 1 & si\ a - w_0 \geq 0 \\ -1 & sinon \end{cases}$$



Le seuil introduit une non-linéarité dans le comportement du neurone, cependant il limite la gamme des réponses possibles à deux valeurs.

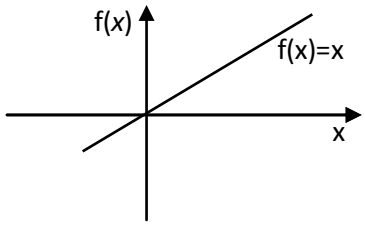
La fonction seuil transforme les signaux d'activation positifs *a* en signaux unité et les signaux d'activation négatifs en signaux nuls. La discontinuité se produit à la valeur d'activation nulle (qui est égal au "seuil" de la fonction signal).

Les fonctions seuils ont été utilisée dans les premiers développements des réseaux de neurones, tels que le perceptron, cependant du fait qu'elles n'étaient pas différentiables, elles ont représentées un obstacle dans le développement des réseaux de neurones jusqu'à l'adoption des fonctions sigmoïdales et le développement des techniques de la descente du gradient (pour l'adaptation des poids).

b) Fonction linéaire :

C'est l'une des fonctions d'activations les plus simples, sa fonction est définie par :

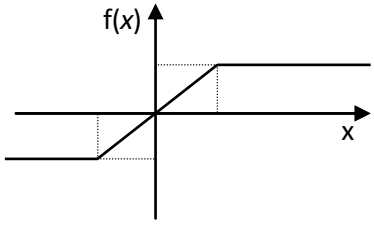
$$f(x) = x$$



c) Fonction linéaire à seuil ou Fonction rampe :

Elle est définie comme suit :

$$f(x) = \begin{cases} x & si\ x \in [u, v] \\ v & si\ x \geq v \\ u & si\ x \leq u \end{cases}$$



Cette fonction représente un compromis entre la fonction linéaire et la fonction seuil : entre ses deux barres de saturation, elle confère au neurone une gamme de réponses possibles. En modulant la pente de la linéarité, on affecte la plage de réponse du neurone.

d) Fonctions non linéaires

Il existe plusieurs types de fonctions non linéaires couramment utilisées comme fonctions d'activation. Ce sont essentiellement les fonctions sigmoïdes et les fonctions gaussiennes, que nous allons voir dans le cadre du neurone non linéaire.

Le type de neurone qui résulte de tous ces genres cités (à seuil, linéaire, à rampe ou sigmoïde) s'appelle neurone sommateur.

Il existe d'autres types de neurones, tels que les neurones distance, les neurones polynomiaux, les neurones de types noyaux … etc.

3. Différentes architectures

La façon dont les neurones d'un réseau sont structurés est intimement liée avec l'algorithme d'apprentissage employé pour entraîner le réseau.

En général, nous pouvons identifier trois classes fondamentalement différentes d'architectures de réseaux :

3.1. Réseau à propagation avant à une seule couche

Dans un réseau de neurones à couches les neurones sont organisés sous forme de couches. Dans la forme la plus simple d'un réseau à couches nous avons une couche d'entrée de nœuds qui se projette sur une couche de sortie (nœuds de calculs), mais pas dans le sens inverse. En d'autres termes, ce réseau est strictement à propagation avant ou de type acyclique. Il est illustré dans la figure 4.2 pour le cas de quatre nœuds dans chacune des couches d'entrée et de sortie. Un tel réseau est appelé réseau à une seule couche, où la désignation '*une seule couche*' se réfère à la couche de sortie composée des nœuds de calcul (neurones). Nous ne comptons pas la couche d'entrée des nœuds source du fait qu'aucun calcul n'y est fait.

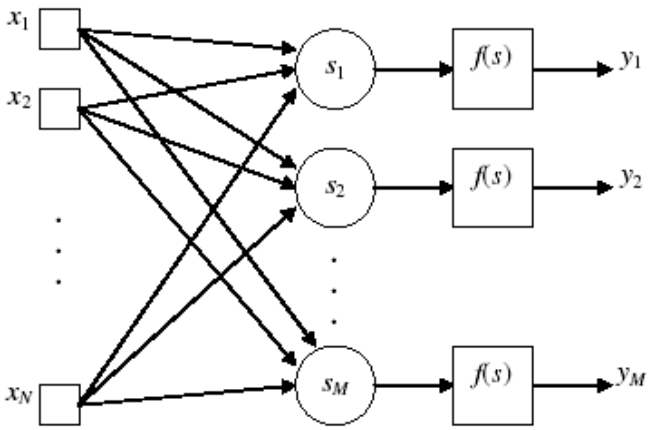


FIGURE 4.2 *: Réseau à propagation avant (ou réseau acyclique) avec une seule couche de neurones.*

3.2. Réseaux multicouches à propagation avant

La seconde classe de réseau à propagation avant se distingue par la présence d'une ou de plusieurs couches cachées, dont les nœuds de calcul sont appelés neurones cachés ou unités cachés. La fonction des neurones cachés est d'intervenir entre l'entrée et la sortie du réseau d'une manière utile. En ajoutant une ou plusieurs couches cachées, le réseau devient capable d'extraire des statistiques d'ordre supérieur. Grosso modo, le réseau acquiert une perspective globale en dépit de sa connectivité locale due à l'ensemble supplémentaire de raccordements synaptiques et à la dimension supplémentaire des interactions neuronales (il y a plus de paramètres et donc plus de degrés de libertés, ce qui donne une plus grande flexibilité).

L'aptitude des neurones cachés d'extraire des statistiques d'ordre supérieur est particulièrement intéressante lorsque la taille de couche d'entrée est grande.

Les nœuds source dans la couche d'entrée du réseau fournissent les éléments de la forme d'activation (le vecteur d'entrée), qui constituent les signaux d'entrée appliqués aux neurones (nœuds de calcul) dans la deuxième couche (i.e., la première couche cachée). Les signaux de sortie de la deuxième couche sont utilisés comme entrée pour la troisième couche, et ainsi de suite pour le reste du réseau. Typiquement, les neurones de chaque couche du réseau ont pour entrées seulement les sorties de la couche précédente. L'ensemble de signaux de sortie des neurones de la dernière couche (finale) du réseau constitue la réponse finale du réseau à la forme d'activation fournie par les nœuds source dans la couche d'entrée (première couche). Le graphe architectural de la figure 4.3 illustre la disposition d'un réseau de neurones multicouches à propagation avant pour le cas d'une seule couche cachée.

Le réseau de la figure 4.3 est dit complètement connecté dans le sens où chaque nœud dans chaque couche est connecté à chaque nœud de la couche suivante. Si cependant, quelques liens de communication (connections synaptiques) manquent au réseau, nous dirons que ce réseau est partiellement connecté.

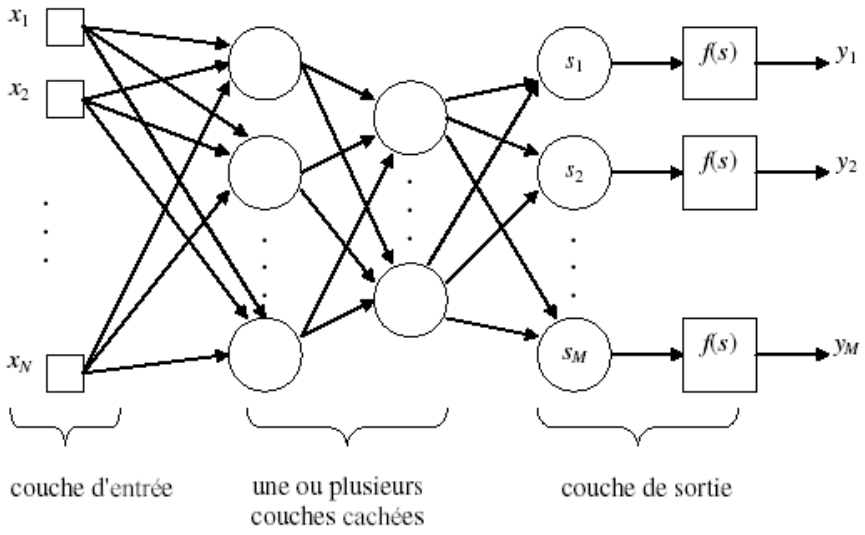


Figure 4.3: réseau à propagation avant (ou acyclique) complètement connecté avec plusieurs couches cachées et une couche de sortie.

QUELQUES RESEAUX MULTICOUCHE

- **LES PERCEPTRONS MULTICOUCHE**

Le réseau consiste en un ensemble d'unités sensorielles qui constituent la couche d'entrée, une ou plusieurs couches cachées, et une couche de sortie.

Le signal d'entrée se propage vers l'avant, de couche en couche.

Les perceptrons multicouches ont été utilisés pour résoudre avec succès des problèmes difficiles et divers, en les entrainant d'une manière supervisée avec l'algorithme de rétro propagation. Cet algorithme est basé sur la règle d'apprentissage de la correction d'erreur.

Le PMC possède trois caractéristiques distinctives:

1. Chaque neurone caché est doté d'une fonction d'activation non linéaire.
2. Le réseau contient une ou plusieurs couches cachées, ce qui permet au réseau d'apprendre des tâches complexes en extrayant progressivement plus de caractéristiques significatives à partir des données.
3. Le réseau présente de hauts degrés de connectivité.

Le perceptron multicouche, entrainé avec l'algorithme de retro propagation, repose sur une forme d'optimisation globale pour sa conception et c'est un approximateur universel

- **RESEAUX A FONCTIONS RADIALES DE BASE (RBF)**

La construction d'un réseau RBF, dans sa forme la plus basique, implique trois couches ayant des rôles entièrement différents. La couche d'entrée est faite d'unités sensorielles qui connectent le réseau à son environnement. La seconde couche, la seule couche cachée du réseau, applique une transformation non linéaire de l'espace des entrées à l'espace caché; dans la majorité des applications l'espace caché est de grande dimension. La couche de sortie est linéaire, fournissant la réponse du réseau au signal d'activation appliqué à la couche d'entrée.

La justification mathématique de la raison de faire suivre une transformation non linéaire par une transformation linéaire réside dans le fait qu'un problème de classification pris dans un espace de grande dimension a plus de chance d'être linéairement séparable que dans un espace de petite dimension.

Un autre point important est le fait que la dimension de l'espace caché est directement liée à la capacité du réseau d'approximer des transformations entrées-sorties lisses. Plus la dimension de l'espace est élevée, plus l'approximation est meilleure.

Dans le PMC, la conception est l'application de 'l'approximation stochastique'. Alors que dans les RBF, l'approche diffère complètement en considérant la conception d'un réseau de neurones comme un problème d'ajustement de courbes dans un espace de grande dimension.

De ce point de vue, apprendre équivaut à trouver une surface dans un espace multidimensionnel qui fournit le meilleur ajustement aux données d'entraînement, où le critère *'meilleur ajustement'* est mesuré dans un certain sens statistique.

La généralisation équivaut à l'utilisation de la surface multidimensionnelle pour interpoler les données de test.

Dans le contexte des réseaux de neurones, les unités cachées fournissent un ensemble de *''fonctions''* qui constituent une *''base*" arbitraire pour les vecteurs d'entrée lorsqu'ils sont exprimés dans l'espace caché; ces fonctions sont appelées *'fonctions radiales de base'*.

Le réseau RBF repose sur une forme d'optimisation locale pour sa conception et c'est un approximateur universel.

- **LES MACHINES A VECTEURS SUPPORTS (SVM) :**

Conséquence de la révolution qui s'est produite en statistique entre les années 60 et les années 80 qui a amené au remplacement du paradigme de Fisher (entre les années 20 et 30) par un nouveau paradigme.

L'apparition de la théorie de Vapnik et de la dimension VC (Vapnik – Chervonenkis) à conduit à la construction de réseau dont le fonctionnement s'appuie sur cette théorie : les machines SVM.

Fondamentalement, le SVM est une machine linéaire possédant d'excellentes propriétés.

Elles permettent de trouver des surfaces discriminantes de n'importe quelle forme, avec un algorithme spécifique.

Un des intérêts des SVM est que la fonction de coût que l'on minimise durant l'apprentissage est convexe (présente un seul minimum), alors la fonction de coût des moindres carrés utilisée pour la régression (ou la fonction de coût d'entropie croisée utilisée pour la classification) présentent des minima locaux.

- Les machines à vecteurs supports exploitent la théorie de la dimension VC pour sa conception et c'est un approximateur universel.

3.3. Réseaux récurrents

Un réseau récurrent se distingue d'un réseau à propagation avant par le fait qu'il possède au moins une boucle arrière. Par exemple, un réseau récurrent peut consister en une seule couche de neurones où chaque neurone fournit son signal de sortie aux entrées de tous les autres neurones, tel qu'illustré dans le graphe architectural de la figure 4.4. Dans la structure représentée dans cette figure il n'y a pas de boucles des neurones vers eux-mêmes ; le self retour réfère à une situation où la sortie du neurone est envoyée à sa propre entrée. Le réseau récurrent de la figure 4.4 ne possède pas de couche cachée.

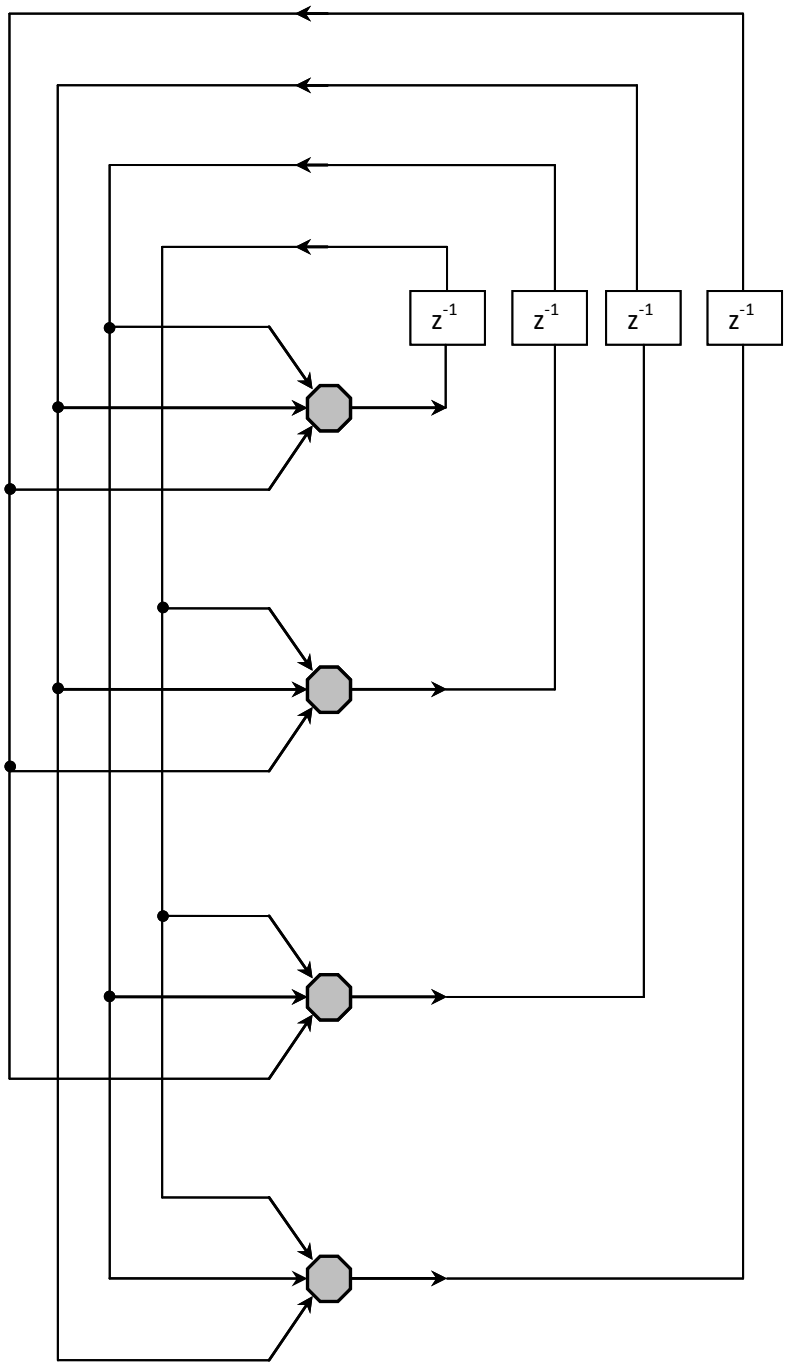


Figure 4.4 : réseau récurrent sans boucle de self retour et sans neurones cachés.

La présence des boucles de retour a un impact profond sur la capacité d'apprentissage du réseau et sa performance. En outre, les boucles de retour impliquent l'utilisation de branches particulières composées d'éléments unités de retard (notés z^{-1}), ce qui se traduit par un comportement dynamique non linéaire (en supposons que le réseau contient des unités non linéaires).

QUELQUES RESEAUX RECURRENTS

- **LE RESEAU DE HOPFIELD**

Le modèle de Hopfield utilise l'architecture des réseaux entièrement connectés et récurrents (dont les connexions sont non orientées et où chaque neurone n'agit pas sur lui-même). Les sorties sont en fonction des entrées et du dernier état pris par le réseau.

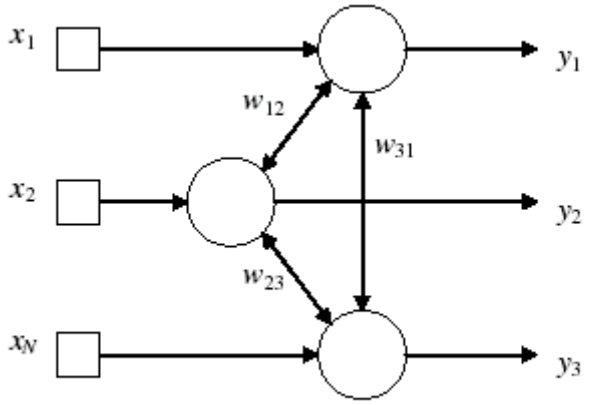


Figure 4.5 : Réseaux associatifs - modèle de Hopfield, réseau récurrent monocouche à connexité totale. Chaque neurone est non seulement l'entrée mais aussi la sortie

C'est un réseau avec des sorties binaires où tous les neurones sont interconnectés avec des poids symétriques, c'est-à-dire que le poids du neurone N_i au neurone N_j est égal au poids du neurone N_j au neurone N_i. Les poids sont donnés par l'utilisateur. Les poids et les états des neurones permettent de définir "l'*énergie*" du réseau. C'est cette énergie que le réseau tente de minimiser pour trouver une solution. La machine de Boltzmann est en principe un réseau de Hopfield, mais qui permet l'apprentissage grâce à la minimisation de cette énergie.

- **LE MODELE DE KOHONEN**

Il a pour objectif de présenter des données complexes et appartenant généralement à un espace discret de grandes dimensions dont la topologie est limitée à une ou deux dimensions. Les cartes de Kohonen sont réalisées à partir d'un réseau à deux couches, une en entrée et une en sortie.

Notons que les neurones de la couche d'entrée sont entièrement connectés à la couche de sortie. (figure 4.6)

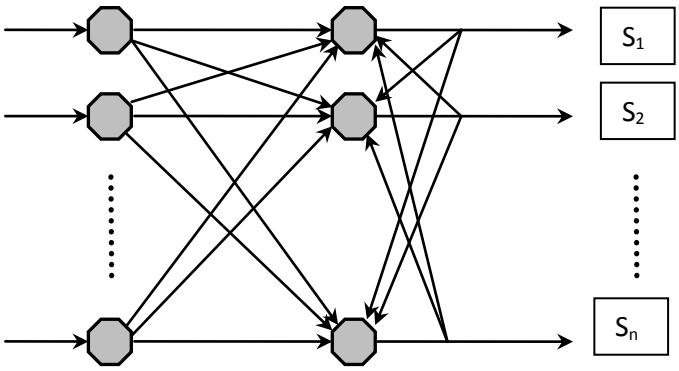


Figure 4.6 : Le modèle de Kohonen

Les neurones de la couche de sortie sont placés dans un espace d'une ou de deux dimensions en général, chaque neurone possède donc des voisins dans cet espace. Et qu'enfin, chaque neurone de la couche de sortie possède des connexions latérales récurrentes dans sa couche.

Une loi d'interaction latérale est aussi modélisée. Les neurones très proches (physiquement) interagissent positivement (le poids des connexions est augmenté autour d'une certaine zone quand une synapse est activée), négativement pour les neurones un peu plus loin, et pas du tout pour les neurones éloignés. Ceci crée un "amas" de neurones activés et contribue à spécialiser certains neurones : pour une entrée donnée, une sortie particulière sera activée alors que les autres resteront inertes. On utilise aussi parfois des lois de concurrence entre les neurones (création et destruction de neurones selon certains critères).

4. Modes d'apprentissage : supervisé et non supervisé

Définition :

L'apprentissage est une phase du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré.

L'apprentissage neuronal fait appel à des exemples de comportement. Le réseau peut ensuite dans une certaine mesure être capable de généraliser, c'est-à-dire de produire des résultats corrects sur des nouveaux cas qui ne lui avaient pas été présentés au cours de l'apprentissage

L'apprentissage est la modification des poids du réseau dans l'optique d'accorder la réponse du réseau aux exemples et à l'expérience

A l'issu de l'apprentissage, les poids sont fixés : c'est alors la phase d'utilisation.

Il existe trois types d'apprentissages principaux. Ce sont l'apprentissage supervisé, l'apprentissage non-supervisé et l'apprentissage par tentative (*graded training en anglais*).

Apprentissage supervisé :

On parle d'apprentissage supervisé quand le réseau est alimenté avec les résultats corrects pour les exemples d'entrées donnés. Le réseau a alors comme but d'approximer ces exemples aussi bien que possible et de développer à la fois la bonne représentation mathématique qui lui permet de généraliser ces exemples pour ensuite traiter de nouvelles situations (qui n'étaient pas présentes dans les exemples).

Après l'apprentissage, le réseau est testé en lui donnant seulement les valeurs d'entrée mais pas les sorties désirées, et en regardant si le résultat obtenu est proche du résultat désiré.

Ainsi, l'apprentissage est dit supervisé lorsque les exemples sont constitués de couples de valeurs du type : (valeur d'entrée, valeur de sortie désirée). Tout le problème de l'apprentissage supervisé consiste, étant donné un ensemble d'apprentissage E de N couples (entrée - sortie désirée) $(x_i, y_i)\ \ i = 1, 2, \dots, n$, à déterminer le vecteur des poids w d'un réseau Fw capable de mettre ces informations en correspondance, c'est à dire un réseau tel que :

$$Fw(x_i) = y_i\ avec\ i = 1, 2, \dots, n$$

Apprentissage non supervisé :

L'apprentissage est qualifié de non supervisé lorsque seules les valeurs d'entrée sont disponibles. Dans ce cas, les exemples présentés à l'entrée provoquent une auto-adaptation du réseau afin d'atteindre une configuration idéale par rapport aux exemples introduits.

Apprentissage par essai - erreur :

C'est un apprentissage où le réseau donne une solution et est seulement alimenté avec une information indiquant si la réponse était correcte ou si elle était au moins meilleure que la dernière fois.

Il existe plusieurs règles d'apprentissage pour chaque type d'apprentissage.

L'apprentissage supervisé est le type d'apprentissage le plus utilisé.

5. Différentes règles d'apprentissage.

5.1. APPRENTISSAGE PAR CORRECTION D'ERREUR

Pour illustrer cette règle d'apprentissage, considérons le cas simple d'un neurone k constitué d'un seul nœud de calcul dans la couche de sortie d'un réseau de neurones à propagation avant, tel que décrit dans la figure 4.7. Le neurone k est dirigé par un vecteur signal **x**(n) produit par une ou plusieurs couches de neurones cachés, qui sont eux-mêmes dirigés par un vecteur d'entrée (stimulus) appliqué aux nœuds sources (i.e., la couche d'entrée) du réseau de neurones. L'argument n dénote le temps discret, ou plus précisément, l'étape de temps d'un processus itératif impliqué dans l'ajustement des poids synaptiques du neurone k. le signal de sortie du neurone k est noté par $y_k(n)$. Ce signal de sortie, représentant la seule sortie du réseau de neurones, est comparé à la réponse désirée ou sortie cible, notée par $d_k(n)$. En conséquence, un signal d'erreur, noté $e_k(n)$, est produit. Par définition, nous avons alors

$$e_k(n) = d_k(n) - y_k(n)$$

Le signal d'erreur $e_k(n)$ enclenche un mécanisme de commande, dont le but est d'appliquer une séquence d'ajustements correctifs aux poids synaptiques du neurone k. les ajustements correctifs sont conçus pour faire que le signal de sortie $y_k(n)$ se rapproche, pas à pas, de la réponse désirée $d_k(n)$. Cet objectif est atteint en minimisant une fonction de coût ou index de performance, $\mathcal{C}(n)$, définie en termes de signal d'erreur $\mathrm{e_k(n)}$ comme :

$$\mathcal{C}(n) = \frac{1}{2}e_k^2(n)$$

Les ajustements pas –à-pas des poids synaptiques du neurone k se poursuivent jusqu'à ce que le système atteigne un état d'équilibre (i.e. les poids synaptiques se stabilisent). En ce point le processus d'apprentissage se termine.

Le processus d'apprentissage décrit ici est appelé *apprentissage par correction d'erreur*. En particulier, la minimisation de la fonction de coût $\mathcal{C}(n)$conduit à une règle d'apprentissage communément appelée *règle delta* ou *règle de Widrow-Hoff*.

Soit $\mathrm{w_{kj}(n)}$ la valeur du poids synaptique $\mathrm{w_{kj}}$ du neurone k excité par l'élément $\mathrm{x_j(n)}$ du vecteur signal **x**(n) à l'étape de temps n. conformément à la règle delta, l'ajustement $\Delta \mathrm{w_{kj}}\,(\mathrm{n})$ appliqué au poids synaptiques $\mathrm{w_{kj}}$ au temps n est défini par

$$\Delta w_{kj}(n) = \eta e_k(n)\mathrm{x_j(n)}$$

Où η est une constante positive qui détermine le taux d'apprentissage lorsque nous passons d'une étape à une autre dans le processus d'apprentissage. Il est alors naturel que nous appelons η *paramètre du taux d'apprentissage*. En d'autres termes, la règle delta peut être définie comme suit :

L'ajustement fait à un poids synaptique d'un neurone est proportionnel au produit du signal d'erreur et du signal d'entrée de la synapse en question.

La règle delta, telle que définie ici, suppose que le signal d'erreur est directement mesurable. Pour que cette mesure soit faisable nous avons clairement besoin que la réponse désirée soit fournie par une source extérieur, qui est directement accessible au neurone k. en d'autres termes, le neurone k est visible au monde extérieur, tel que décrit par la figure 4.7.
L'apprentissage par correction d'erreur est local dans sa nature. C'est-à-dire que les ajustements synaptiques faits par la règle delta sont localisés autour du neurone k.

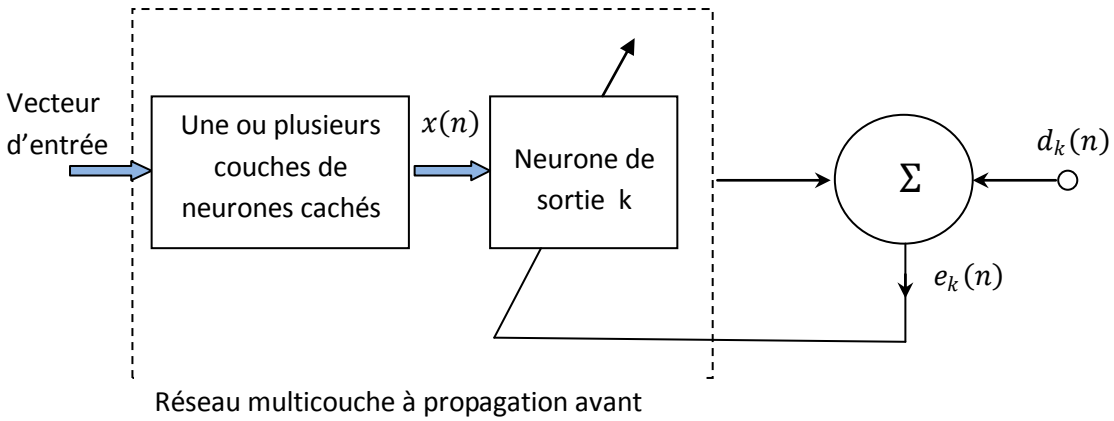


Figure 4.7 : schéma fonctionnel d'un réseau de neurones ayant un seul neurone de sortie

Ayant calculé l'ajustement synaptique $\Delta w_{kj}\,(n)$, la valeur mise à jour du poids synaptique w_{kj} est déterminée par

$$w_{kj}\,(n+1) = w_{kj}\,(n) + \Delta w_{kj}\,(n)$$

En effet, $w_{kj}\,(n)$ et $w_{kj}\,(n+1)$ peuvent être vues respectivement comme l'ancienne et la nouvelle valeur du poids synaptique w_{kj}.

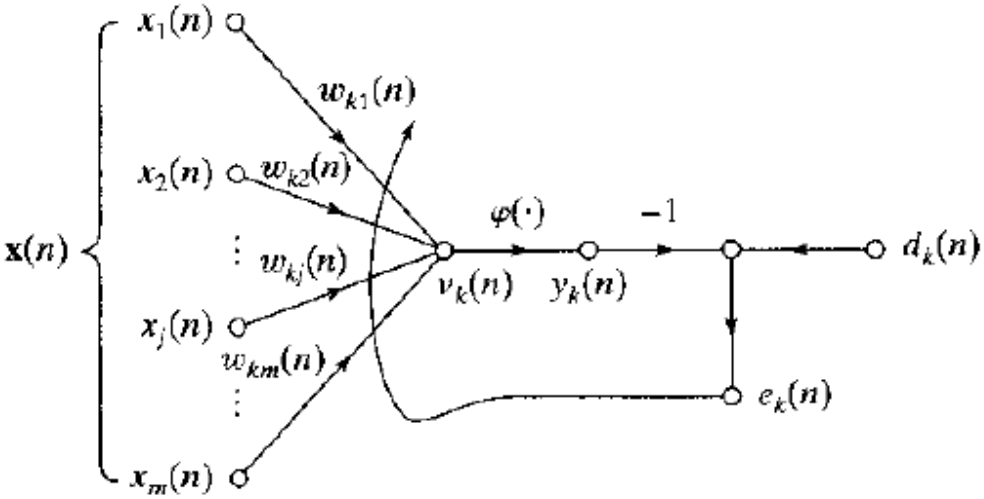


Figure 4.8 : graphique du flux de signal du neurone de sortie

La figure 4.8 montre une représentation du graphique du flux de signal du processus d'apprentissage par correction d'erreur, focalisant sur l'activité entourant le neurone k. Le signal d'entrée $\mathrm{x_j}$ et le champ local induit (le potentiel) $\mathrm{v_k}$ du neurone k sont respectivement appelé signaux pré-synaptique et post-synaptique de la $\mathrm{j^{ème}}$ synapse du neurone k.

Il est important que η soit choisi avec précaution pour assure la stabilité ou la convergence du processus itératif d'apprentissage. Le choix de η a aussi une profonde influence sur l'exactitude et d'autre aspects du processus d'apprentissage. Le paramètre su taux d'apprentissage joue un rôle clé dans la performance en pratique de l'apprentissage par correction d'erreur.

5.2. APPRENTISSAGE BASE SUR LA MEMOIRE

Dans l'apprentissage basé sur la mémoire, tout (ou la majorité) des expériences passées sont explicitement stockées dans une grande mémoire d'exemples d'entrées-sorties correctement classés : $\{(x_i, d_i)\}_{i=1}^{N}$, où $\mathbf{x}_\mathrm{i}$ représente un vecteur entrée et $\mathrm{d_i}$ représente la sortie désirée. Sans perte de généralité, nous pouvons considérer que la réponse désirée soit un scalaire. Par exemple, dans un problème de classification binaire il y a deux classes (hypothèses), notées C_1 et C_2. Dans cet exemple, la réponse désirée $\mathrm{d_i}$prend la valeur 0 pour la classe C_1 et la valeur 1 pour la classe C_2. Lorsque la classification d'un vecteur test $\mathbf{x}_{\text{test}}$ (qui n'est pas vu auparavant) est requise, l'algorithme répond en cherchant et en analysant les données d'entraînement dans un ''voisinage local'' de $\mathbf{x}_{\text{test}}$.

Tous les algorithmes d'apprentissage basé sur la mémoire impliquent deux ingrédients essentiels :

- Un critère utilisé pour définir le voisinage local du vecteur test $\mathbf{x}_{\text{test}}$.
- Une règle d'apprentissage appliquée aux exemples d'entraînement dans le voisinage local de $\mathbf{x}_{\text{test}}$.

Les algorithmes diffèrent les uns des autres par la manière dont ces deux ingrédients dont définis.

Dans un type, simple mais efficace, d'apprentissage basé sur la mémoire connu sous le nom de règle du plus proche voisin, le voisinage local est défini comme étant l'exemple d'apprentissage qui se situe dans le voisinage immédiat du vecteur test $\mathbf{x}_{\text{test}}$. En particulier, le vecteur

$$\boldsymbol{x}_N^{'} \in \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$$

Est dit le plus proche voisin de $\mathbf{x}_{\text{test}}$ si

$$\min_i d(\boldsymbol{x}_i, \mathbf{x}_{\text{test}}) = d(\boldsymbol{x}_N^{'}, \mathbf{x}_{\text{test}})$$

Où $d(\boldsymbol{x}_N^{'}, \mathbf{x}_{\text{test}})$ est la distance euclidienne entre les vecteurs $\boldsymbol{x}_i$ et $\mathbf{x}_{\text{test}}$. La classe associée à ce vecteur $\boldsymbol{x}_N^{'}$ ayant la distance minimale est celle où sera classé $\mathbf{x}_{\text{test}}$.

Une variante du classificateur par plus proche voisin est le classificateur par les k plus proches voisins, qui procède comme suit :

- Identifier les k vecteurs classés qui soient les plus proches du vecteur test $\mathbf{x}_{\text{test}}$.
- Assigner $\mathbf{x}_{\text{test}}$ à la classe qui est la plus fréquemment représentée parmi ces k plus proches voisins (i.e., en utilisant un vote majoritaire pour faire la classification).

En particulier, ce classificateur discrimine contre une valeur aberrante.

5.3. APPRENTISSAGE DE HEBB

L'apprentissage de Hebb repose sur la formulation suivante :

1. *Si deux neurones sur les deux côtés d'une synapse (connexion) sont activés simultanément (i.e., d'une manière synchrone), alors la force de cette synapse est sélectivement augmentée.*
2. *Si deux neurones sur les deux côtés d'une synapse sont activés asynchroniquement, alors la synapse est sélectivement affaiblie ou éliminée.*

Une telle synapse est appelée *synapse hebbienne*. (la règle originale de Hebb ne contient pas la partie 2). Plus précisément, une synapse hebbienne est définie comme une synapse *qui utilise un mécanisme dépendant du temps, hautement local, et fortement interactif pour augmenter l'efficacité de la synapse en tant que fonction de la corrélation entre les activités pré synaptique et post synaptique*.

Modèles mathématiques des modifications hebbiennes

Afin de formuler en termes mathématiques l'apprentissage hebbien, considérons le poids synaptiques $\mathrm{w_{kj}}$ du neurone k avec les signaux pré et post synaptiques notés respectivement par $\mathrm{x_j}$ et $\mathrm{y_k}$. l'ajustement appliqué au poids $\mathrm{w_{kj}}$ à l'instant n est exprimé dans la forme générale

$$\Delta w_{kj}(n) = F\left(y_k(n), x_j(n)\right)$$

Où $F(.,.)$ est une fonction des signaux pré synaptique et post synaptique. Les signaux $\mathrm{x_j(n)}$ et $\mathrm{y_k(n)}$sont souvent considérés sans dimension. La formule de l'équation précédente admet beaucoup de formes, elles sont toutes qualifiées comme hebbienne.

Nous allons considérer deux de ces formes.

Hypothèse de Hebb

La forme la plus simple de l'apprentissage hebbien est décrite par

$$\Delta w_{kj}(n) = \eta y_k(n)\mathrm{x_j(n)} \quad (4.1)$$

Où η est une constante positive qui détermine le taux d'apprentissage. L'équation précédente montre clairement la nature corrélationnelle de la synapse hebbienne. Elle est quelquefois appelée la règle du produit d'activités. La courbe du haut de la figure 4.9 montre une représentation graphique de l'équation (4.1) où le changement $\Delta\mathrm{w_{kj}}$ est représenté en

fonction du signal de sortie (activité post synaptique) y_k. Nous voyons à partir de cette représentation qu'une application répétée du signal d'entrée (activité pré synaptique) x_j conduit à une augmentation de y_k et qui donc augmente d'une manière exponentielle ce qui conduira finalement la connexion synaptique à la saturation. En ce point aucune information ne peut désormais être stockée dans la synapse et ainsi la sélectivité ne peut plus se faire.

Hypothèse de la covariance

Un moyen de surmonter la limitation de l'hypothèse de Hebb est d'utiliser l'hypothèse de la covariance introduite dans Sejnowski (1977). dans cette hypothèse, les signaux pré et post synaptiques sont remplacés dans l'équation (4.1) par leurs différences à leurs valeurs moyennes sur un certain intervalle de temps. Soit $\bar{x}$ et $\bar{y}$ les moyennes respectives sur le temps des signaux présynaptique x_j et post synaptique y_k.conformément à l'hypothèse de la covariance, l'ajustement appliqué au poids synaptiques w_{kj} est défini par

$$\Delta w_{kj} = \eta(x_j - \bar{x})(y_k - \bar{y}) \qquad (4.2)$$

Où η est le paramètre du taux d'apprentissage. Les valeurs moyennes $\bar{x}$ et$\bar{y}$ constituent les seuils pré synaptique et post synaptique, qui déterminent le signe de la modification synaptique. En particulier, l'hypothèse de la covariance conduit à ce qui suit :

- Convergence à un état non trivial, qui est atteint lorsque $x_j = \bar{x}$ ou $y_k = \bar{y}$.
- Prédiction de la potentiation synaptique (i.e., augmentation de la force synaptique) et la dépression synaptique (i.e., diminution de la force synaptique).

La figure 4.9 illustre la différence entre l'hypothèse de Hebb et l'hypothèse de la covariance.

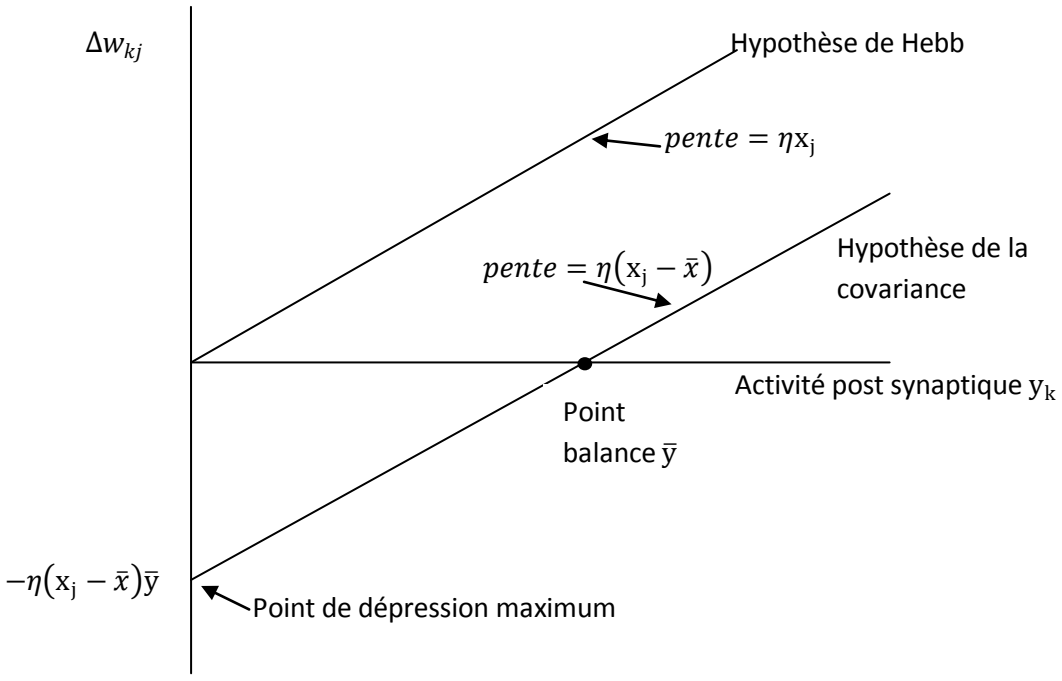


Figure 4.9 : illustration de l'hypothèse de Hebb et de l'hypothèse de la covariance.

Dans les deux cas la dépendance de Δw_{kj} de y_k est linéaire ; cependant, l'interception de l'axe de y_k dans l'hypothèse de Hebb est à l'origine, alors que dans l'hypothèse de la covariance elle est à $y_k = \bar{y}$.

Nous pouvons faire, à partir de l'équation (4.2) les importantes observations suivantes :

1. Le poids synaptique w_{kj} est amélioré s'il y a des niveaux d'activité pré et post synaptiques suffisants, c'est-à-dire que les conditions $x_j > \bar{x}$ et $y_k > \bar{y}$ sont toutes les deux satisfaites.
2. Le poids synaptique w_{kj} est diminué si
    - Il y a une activation pré synaptique (i.e., $x_j > \bar{x}$) en l'absence d'une activation post synaptique suffisante (i.e., $y_k < \bar{y}$).
    - Il y a une activation post synaptique (i.e., $y_k > \bar{y}$) en l'absence d'une activation pré synaptique suffisante (i.e., $x_j < \bar{x}$).

Ce comportement peut être vu comme une compétition temporelle entre les formes d'entrées.

5.4. APPRENTISSAGE COMPETITIF

Dans l'apprentissage compétitif, comme son nom l'indique, les neurones de sortie du réseau de neurones concourent entre eux pour devenir actifs. Tandis que dans un réseau de neurones basé sur l'apprentissage de Hebb plusieurs neurones de sortie peuvent être actifs simultanément, dans l'apprentissage compétitif un seul neurone est actif à la fois. C'est cette caractéristique qui fait que l'apprentissage compétitif est fortement adapté pour découvrir des traits statistiques saillants qui peuvent être utilisés pour classer un ensemble de données.

Il y a trois éléments de base pour une règle d'apprentissage compétitif :

- Un ensemble de neurones qui sont tous les mêmes sauf pour quelques poids synaptiques aléatoirement distribués, et qui donc répondent différemment à un ensemble donné de formes d'entrée.
- Une limite imposée à la ''force'' de chaque neurone.
- Un mécanisme qui permet aux neurones de concourir pour l'exclusivité à répondre à un sous ensemble donné d'entrées, de sorte que seul un neurone de sortie, ou seul un neurone par groupe, est actif à la fois. Le neurone qui gagne la compétition est appelé un *neurone tout gagnant*.

De cette manière, les neurones du réseau apprennent à se spécialiser sur des ensembles de formes similaires ; et de cette manière ils deviennent des détecteurs de caractéristiques pour des classes différentes de formes d'entrée.

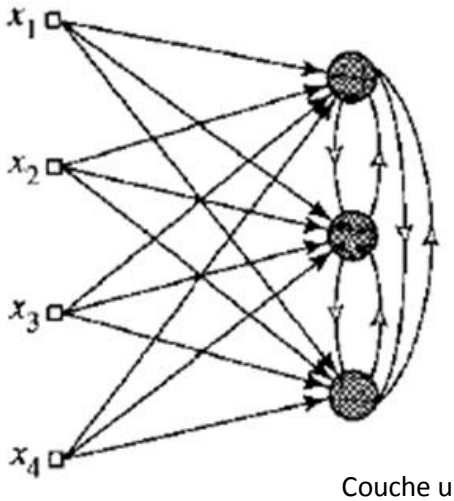


Figure 4.10 : graphe architectural d'un réseau simple à apprentissage compétitif avec des connections à propagation avant (excitatrices) à partir des nœuds sources, et des connections latérales (inhibitrices).

Dans la forme la plus simple de l'apprentissage compétitif, le réseau de neurones possède une seule couche de neurones de sortie, chacun est entièrement connecté aux nœuds sources. Le réseau peut inclure des connections de retour, comme indiqué dans la figure précédente. Dans le réseau décrit ici, les connections produisent des inhibitions latérales, où chaque neurone tend à inhiber le neurone auquel il est latéralement connecté. Par contre, les connections synaptiques à propagation avant, dans le réseau de cette figure, sont toutes excitatrices.

Pour qu'un neurone k soit le neurone vainqueur, son champ induit local v_k pour une entrée spécifiée **x,** doit être le plus grand parmi les neurones du réseau. Le signal de sortie y_k du neurone vainqueur k est posé égal à 1 ; les signaux de sortie de tous les neurones qui ont perdu la compétition sont posés égaux à 0. Nous écrivons donc

$$y_k = \begin{cases} 1 & si\ v_k > v_j \quad pour\ tout\ j, \qquad j \neq k \\ 0 & sinon \end{cases}$$

Où le champ local induit v_k représente l'action combinée de toutes les entrées, avant et arrière du neurone k.

Soit w_{kj} le poids synaptique connectant le neurone d'entrée j au neurone k. supposons que chaque neurone est doté d'une quantité fixée de poids synaptique (i.e., tous les poids synaptiques sont positifs), qui sont distribués sur les nœuds d'entrée ; c'est-à-dire,

$$\sum_j w_{kj} = 1 \ \ pour\ tout\ k$$

Alors, un neurone apprend en décalant les poids synaptiques de ses nœuds d'entrée inactifs aux nœuds actifs. Si un neurone ne répond pas à une forme d'entrée particulière, le neurone

n'enregistre aucun apprentissage. Si un neurone particulier gagne la compétition, chaque nœud d'entrée abandonne une certaine proportion de son poids synaptique, et le poids abandonné est alors distribué sur les nœuds d'entrée actifs. Conformément à *la règle standard de l'apprentissage compétitif*, le changement Δw_{kj} appliqué au poids synaptique w_{kj} est défini par

$$\Delta w_{kj} = \begin{cases} \eta(x_j - w_{kj}) & si\ le\ neurone\ k\ gagne\ la\ compétition \\ 0 & si\ le\ neurone\ k\ perd\ la\ compétition \end{cases}$$

Où η est le paramètre taux d'apprentissage. Cette règle provoque un effet global de mouvement du vecteur des poids synaptiques $\mathbf{w}_k$ du neurone gagnant k en direction de la forme d'entrée **x**.

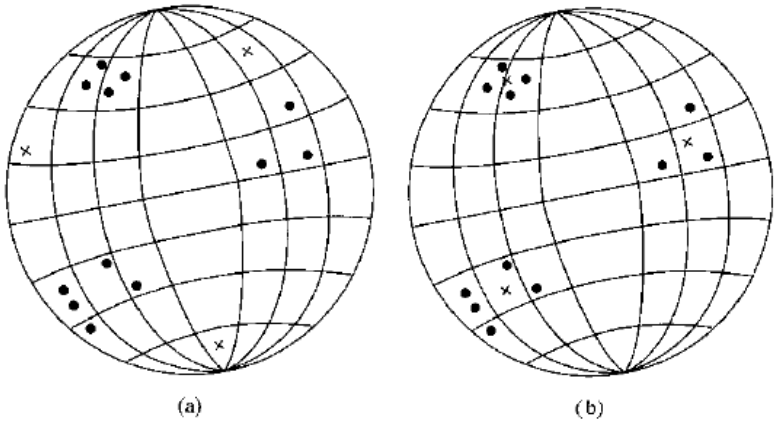


Figure 4.11 : interprétation géométrique du processus de l'apprentissage compétitif. Les points représentent les vecteurs entrée, les croix représentent les vecteurs des poids synaptiques de trois neurones de sortie.

(a) L'état initial du réseau
(b) L'état final du réseau

Nous allons utiliser l'analogie géométrique décrite dans la figure précédente pour illustrer l'essence de l'apprentissage compétitif. Il est supposé que chaque vecteur entrée **x** a une longueur euclidienne constante de sorte que nous pouvons le voir comme un point sur une sphère N-dimensionnelle où N est le nombre de nœuds sources. N représente aussi la dimension de chaque vecteur de poids synaptiques $\mathbf{w}_k$. il est en plus supposé que tous les neurones du réseau sont contraints d'avoir le même longueur euclidienne (norme), telle que montrée par

$$\sum_j w_{kj}^2 = 1 \ \ pour\ tout\ k$$

Quand les poids sont correctement mesurés ils forment un ensemble de vecteurs qui tombent sur la même sphère unité de dimension N. Dans la figure 4.11a nous montrons trois groupements naturels de formes (stimulus) représentés par des points. Cette figure inclus aussi un état initial possible du réseau (représenté par des croix) qui peut exister avant l'apprentissage. La figure 4.11b montre un état final typique du réseau qui résulte de l'utilisation de l'apprentissage compétitif. En particulier, chaque neurone de sortie a découvert un groupe de formes d'entrée en déplaçant son

vecteur de poids synaptiques sur le centre de gravité de ce groupe découvert. Cette figure illustre la capacité du réseau de neurones de produire une agrégation au moyen de l'apprentissage compétitif.

Cependant, pour que cette fonction soit faite d'une manière ''stable'' les formes d'entrée doivent tomber dans des groupements suffisamment distincts avec lesquels on peut commencer. Sinon le réseau peut être instable par le fait qu'il ne peut répondre à deux formes appartenant à deux classes empiètentes avec le même neurone de sortie.

5.5. APPRENTISSAGE DE BOLTZMANN

L'apprentissage de Boltzmann est un algorithme d'apprentissage stochastique déduit d'idées provenant de la mécanique statistique. Un réseau de neurone conçu sur la base de la règle d'apprentissage de Boltzmann est appelé machine de Boltzmann.

Dans une machine de Boltzmann les neurones constituent une structure récurrente, et ils opèrent d'une manière binaire, dans le sens où ils sont soit dans l'état actif, noté par +1, soit dans l'état passif, noté par -1. La machine est caractérisée par une *fonction d'énergie,* E, dont la valeur est déterminée par les états particuliers pris par les neurones de la machine. Elle s'exprime de la manière suivante

$$E = -\frac{1}{2}\sum_{j}\sum_{k} w_{kj}\, x_k x_j \qquad j \neq k$$

Où x_j est l'état du neurone *j*, et w_{kj} est le poids de la connexion reliant le neurone j au neurone *k*. le fait que $j \neq k$ signifie qu'il n'y a pas de neurone dans la machine qui ait un retour vers lui-même. La machine opère en choisissant aléatoirement un neurone *k* en une certaine étape du processus d'apprentissage ; elle bascule ensuite le neurone *k* de l'état x_k à l'état $-x_k$ à une certaine température T avec la probabilité

$$P(x_k \rightarrow -x_k) = \frac{1}{1 + exp(-\Delta E_k / T)}$$

Où ΔE_k est le changement d'énergie (c'est-à-dire le changement dans la fonction d'énergie de la machine) résultant d'un tel basculement. Il faut noter que *T* n'est pas la température physique, mais plutôt une *pseudo température.* Lorsque la règle est appliquée d'une manière répétée, la machine va atteindre un *équilibre thermique.*

Les neurones de la machine de Boltzmann se répartissent en deux groupes fonctionnels : visibles et cachés. Les neurones visibles fournissent une interface entre le réseau et l'environnement dans lequel il opère, tandis que les neurones cachés opèrent toujours librement.

Il y a deux modes d'opérations qui doivent être considérés :

- *Condition fixée,* dans laquelle les neurones visibles sont fixés sur des états spécifiques déterminés par l'environnement.
- *Condition libre,* dans laquelle tous les neurones (visibles et cachés) opèrent librement.

Soit ρ_{kj}^{+} la corrélation entre les états du neurone j et du neurone k, lorsque le réseau est dans la condition fixée. Soit ρ_{kj}^{-} la corrélation entre les états du neurone j et du neurone k, lorsque le réseau est dans la condition libre. Les deux corrélations sont calculées comme moyenne sur tous les états possibles de la machine lorsqu'elle est dans l'équilibre thermique. Alors, conformément à la règle de Boltzmann, le changement Δw_{kj} appliqué au poids synaptiques w_{kj} est défini par

$$\Delta w_{kj} = \eta(\rho_{kj}^{+} - \rho_{kj}^{-}), \quad j \neq k$$

Où η est le taux d'apprentissage.

6. Les réseaux de neurones dans la reconnaissance des formes non linéaires

Il y a plusieurs méthodes appropriées pour l'analyse non linéaire, parmi elles figurent les réseaux perceptron multicouche (MLP), les réseaux à fonctions radiales de base (RBF), les machines à vecteur support (SVM), les modèles généralisés de traitement des données (GMDH), qui sont aussi appelés les réseaux polynômiaux, les réseaux de neurones pour la régression généralisée (GRNN) et les réseaux de neurones généralisés (GNN). La plupart de ces réseaux ont plusieurs couches de traitement qui leur donnent la capacité de modélisation non linéaires.

- MLP : réseaux perceptron multicouche : utilisent une grande variété de fonctions d'activation
- RBF : réseaux à fonctions radiales de base : utilise les fonctions gaussiennes comme fonction d'activation
- SVM : machines à vecteurs supports : s'appuient sur une nouvelle méthode statistique qui transforme les problèmes multidimensionnels non linéaires en problèmes linéaires dans des espaces de plus grandes dimensions.
- GMDH : modèles généralisés pour le traitement des données (generalized model for data handling) qui sont aussi appelés les réseaux polynômiaux : c'est un réseau non linéaire qui construit des polynômes à partir des variables d'entrée dans des étapes successives afin d'obtenir des polynômes plus complexes contenant les variables d'entrée les plus pertinentes.
- GRNN : réseau de neurones pour la régression généralisée (generalized regression neural network) : il est utilisé pour l'estimation non paramétrique des densités de probabilités des données ; il ne requiert pas un entraînement itératif et il est utilisé pour le traitement de données relativement non linéaires.
- GNN : réseau de neurone généralisé (generalized neural network) : c'est un réseau qui a seulement deux neurones cachés, l'un qui exécute une analyse linéaire et l'autre qui exécute une analyse non linéaire.

Parmi les types de réseaux cités ci-dessus, le plus largement utilisé est le perceptron multicouche.

La configuration d'un réseau perceptron multicouche est représentée dans la figure ci-dessous.

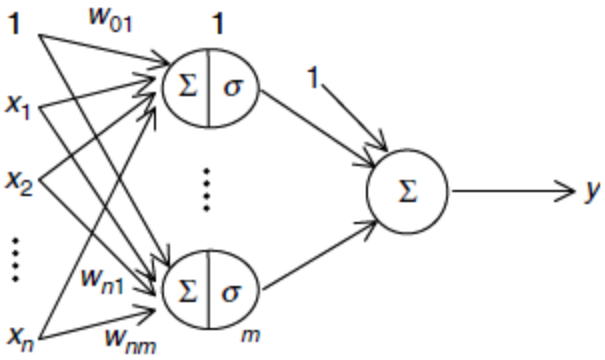


Figure 4.12

Dans cette figure, $x_1, x_2, \dots, x_n$ sont les variables d'entrée et qui constituent la couche d'entrée. Le premier ensemble de flèches représente les poids (ou les connections entrées –neurones cachés) qui lient cette couche à la couche cachée intérieure, constituée d'un ou de plusieurs neurones cachés. Ils sont ainsi appelés car ils ne sont pas exposés à l'environnement extérieur (données), comme le sont les neurones d'entrée et de sortie. Les neurones cachés somment les entrées pondérés tel que noté par le symbole $\sum$ dans la figure 4.12 ; ce qui est similaire au premier traitement dans le neurone linéaire et dans le perceptron. Cependant, contrairement à ces systèmes, chaque neurone caché fait passer sa somme pondérée à travers une fonction de transfert non linéaire, notée par σ. Les sorties de ces neurones cachés, à travers les connections neurones cachés-neurones de sortie, alimentent les entrées du neurone (ou des neurones) de sortie qui réuni les sorties en calculant la somme pondérée et en la faisant passer à travers une fonction qui peut être linéaire ou non linéaire. Les sorties de ces neurones constituent les sorties du réseau. Ce qui, habituellement, est une seule sortie dans le cas de la prédiction (ou de l'approximation de fonction), et une ou plusieurs sorties dans le cas de la classification, pour indiquer les classes auxquelles devront appartenir les observations.

Il y a plusieurs choix pour la fonction d'activation (ou de transfert), et nous allons exposer les raisons de ces choix, de leurs caractéristiques, et la manière dont elles transforment les réseaux multicouche en classificateurs et prédicteurs puissants.

L'apprentissage ou l'entraînement dans les MLP est supervisé dans le sens où les réponses correctes sont présentées au réseau pour chacune des entrées.

6.1. Les neurones non linéaires

Les MLP tiennent leur puissance du traitement non linéaire effectué dans les neurones cachés. Ce sont les fonctions d'activation non linéaires qui sont cruciales pour cette tâche en transformant d'une manière non linéaire l'entrée pondérée du neurone en sortie.

Nous allons nous intéresser en détail à quelques unes de ces fonctions d'activations non linéaires, puis elle nous allons étudier comment elles sont transformées elles mêmes durant le processus où elles s'entraînent à projeter d'une manière non linéaire les entrées en sortie(s).

La figue 4.13 montre un neurone caché recevant n entrées, $x_1, x_2, \dots, x_n$.

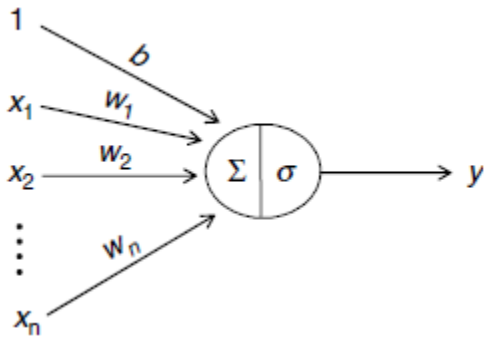


Figure 4.13 : Neurone non linéaire

La sortie du neurone est donnée par

$$\sigma\left(\sum_{j=1}^{n} w_j x_j + b\right)$$

Où la somme pondérée des entrées (entre parenthèses) est passée à travers une fonction non linéaire σ. b représente le poids associé au biais et w_j représente le poids associé à la $j^{ème}$ entrée.
La fonction la plus communément utilisée est la fonction sigmoïde – c'est une famille de courbes incluant la fonction logistique et la fonction tangente hyperbolique. Elles sont utilisées dans la modélisation de la dynamique des populations, dans les sciences économiques et dans d'autres domaines.
Les autres fonctions utilisées sont la fonction gaussienne, la fonction sinus, la fonction arc tangente, et leurs variantes.

6.2. Fonctions d'activation de neurones

Les fonctions d'activation montrées dans la figure suivante possèdent des caractéristiques importantes qui les rendent vitales pour le traitement de l'information neuronale.

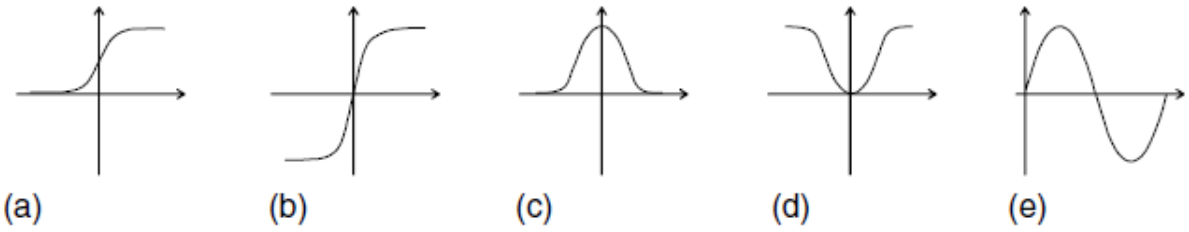


Figure 4.14

a) Fonction logistique
b) Tangente hyperbolique
c) Fonction gaussienne
d) Fonction gaussienne inverse
e) Fonction sinus

Elles sont **non linéaires**, **continues** et **bornées**.

Non linéaire signifie que la sortie de la fonction varie d'une manière non linéaire avec l'entrée ; cet aspect rend possible pour les réseaux de neurones de produire des applications non linéaires entre les entrées et les sorties.

La continuité de la fonction implique qu'il n'y a pas de coupures et c'est une condition nécessaire pour que la fonction soit différentiable, c'est ce qui rend possible la mise en œuvre de la règle delta pour ajuster les poids par la rétro propagation des erreurs.

Ces deux importantes propriétés seront utilisées pour étendre les domaines d'application des réseaux de neurones du simple domaine linéaire au complexe domaine non linéaire.

Le terme **« bornées »** signifie que la sortie ne peut jamais atteindre de trop grandes valeurs, quelque soient les entrées.

Ces développements ont été le résultat directe des tentatives pour développer des modèles qui miment les neurones biologiques où les sorties sont des signaux non linéaires, continues et bornés.

Les fonctions d'activation sigmoïdes sont les plus populaires. Les concepts évoqués ci-dessus s'appliquent tout aussi bien aux autres fonctions.

6.2.1. Les fonctions sigmoïdes

Les fonctions sigmoïdes sont une famille de fonctions en forme de S ; deux d'entre elles sont représentées dans les figures 4.14a et 4.14b. la fonction sigmoïde la plus utilisée est la fonction logistique. Voir la figure ci-dessous où elle est représentée pour des valeurs de u allant de -10 à 10. L(u) représente la sortie pour un input u.

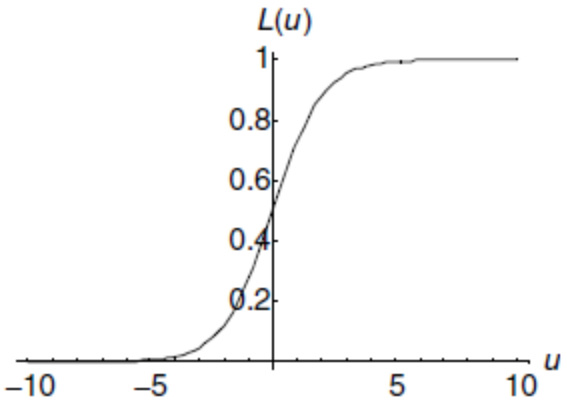


Figure 4.15

La borne inférieure de la fonction logistique est zéro, et la borne supérieure est 1. Ce qui signifie que l'étendue de la fonction est $[0\ ,1]$.

Au point $u = 0$ la sortie est le point moyen (0.5), et la pente de la fonction, qui indique à quelle vitesse la fonction change, est le plus grand en ce point.

La pente en $u = 0$ est 0.25 (14°).

La sortie croît relativement rapidement dans le voisinage de $u = 0$, mais croît beaucoup plus lentement à l'approche de la borne supérieure.

Pour les entrées inférieures à zéro, la sortie décroît d'abord rapidement puis plus lentement à l'approche de la borne inférieure.

La formule de la fonction logistique est la suivante :

$$y = L(u) = \frac{1}{1 + e^{-u}}$$

Où e est le nombre d'Euler.

Une autre fonction sigmoïde communément utilisée est la tangente hyperbolique montrée dans la figure 4.14b et donnée par la formule :

$$\tanh(u) = \frac{1 + e^{-u}}{1 - e^{-u}}$$

La figure suivante montre que la fonction tangente hyperbolique a une borne inférieure égale à -1 et une borne supérieure égale à 1. Ce qui fait que son étendue est $[-1\ ,1]$ contrairement à celui de la fonction logistique où il est $[0\ ,1]$.

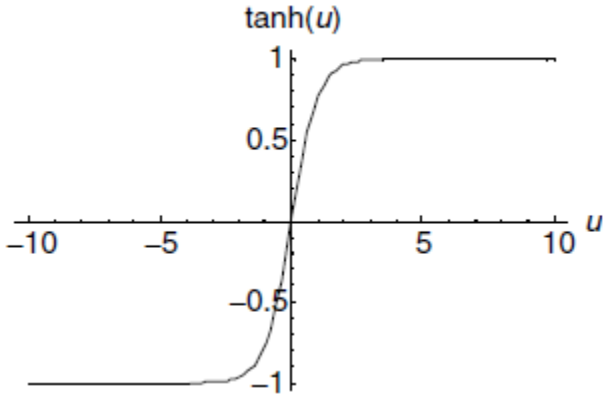


Figure 4.16

Une autre différence est que la sortie au point $u = 0$ est zéro.

De même, la pente en ce point $u = 0$ est plus grande, ce qui signifie que la tangent hyperbolique atteint ses bornes plus rapidement que la fonction logistique.

La pente au point $u = 0$ est égale à 1.0 (i.e., 45°).

Une autre fonction apparentée est l'inverse de la tangent (ou arc tangente), montrée dans la figure ci-dessous :

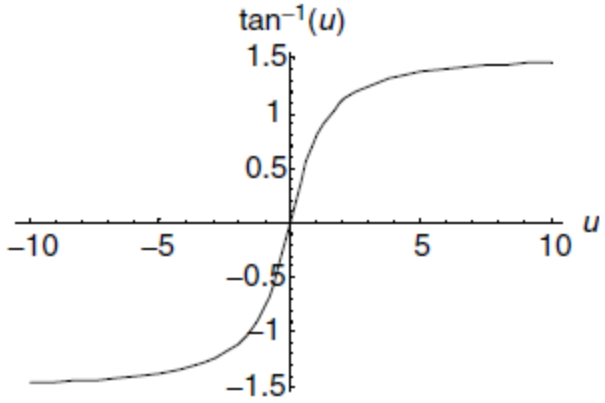


Figure 4.17 : fonction arctangente

Cette fonction a une variation plus graduée que les deux fonctions précédentes, avec une pente intermédiaire entre ceux de la fonction logistique et la fonction tangente hyperbolique.

Plusieurs de ces fonctions ont été utilisées pour le traitement de l'information neuronale. Lorsqu'elles sont utilisées dans les neurones de sortie, l'étendue de la sortie cible doit coïncider avec l'étendue de la fonction de transfert car la sortie du réseau doit être comparée à cette sortie cible. Par exemple, si la fonction logistique est utilisée dans le neurone de sortie, le domaine de la sortie sera l'intervalle $[0\ ,1]$ et donc la sortie cible doit être adaptée pour entrer dans cet intervalle.

Les entrées, cependant, peuvent prendre n'importe quelle valeur, indépendamment des bornes de la fonction sigmoïde.

6.2.2. Les fonctions gaussiennes

- *La courbe normale standard*

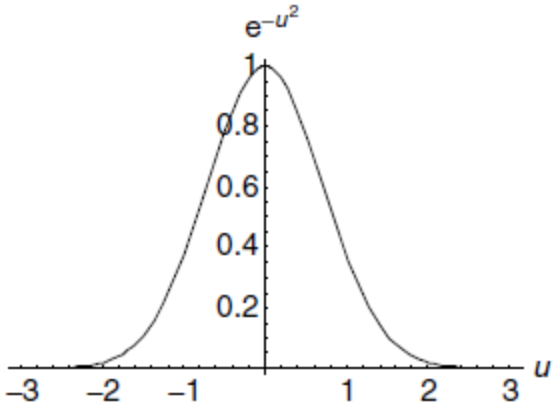


Figure 4.18

La courbe normale standard a la forme d'une cloche symétrique et son équation est

$$y = e^{-u^2}$$

Sa moyenne est égale à zéro et son écart type est égal à 1.

Son étendue est $[0\ ,1]$ et elle atteint son maximum au point $u = 0$.

Elle est hautement sensible pour les valeurs de u proche de zéro, et elle est presque insensible à celles sur les queues. Elle amplifie donc la partie intérieure de la distribution des données.

De ce fait, le neurone dans lequel est utilisé cette fonction est plus sensible aux entrées pondérées voisines de zéro.

- *La fonction complémentaire de la gaussienne*

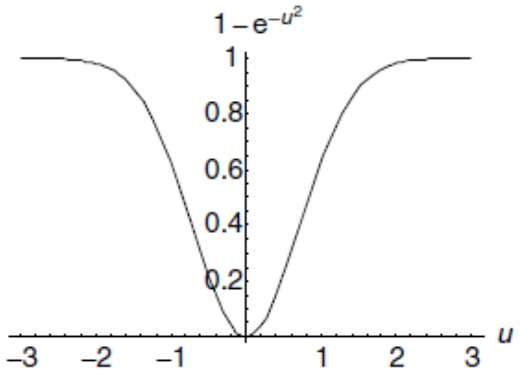


Figure 4.19

Son équation est

$$y = 1 - e^{-u^2}$$

C'est l'inverse de la fonction gaussienne, et donc elle est maximale sur les bord et elle vaut zéro quand $u = 0$.

Quand elle est utilisée dans un neurone faisant partie d'un réseau, le réseau devient plus sensible aux données qui se trouvent aux extrémités.

Remarques :

Fonctionnellement, la performance d'un seul neurone non linéaire est similaire à la régression non linéaire en statistique.

En théorie, nous pouvons utiliser n'importe quelle fonction d'activation, cependant, il faut noter qu'une fonction d'activation linéaire dans le neurone de sortie est plus appropriée pour la prédiction et la fonction logistique ou une fonction bornée est plus appropriée pour la classification.

Cas particulier

Réseaux à une couche cachée de sigmoïdes et un neurone de sortie linéaire

C'est des réseaux qui ont une importance particulière dans la prédiction et l'approximation des fonctions, en vertu de la propriété qui va suivre.

La figure 4.20 représente un réseau à une couche cachée à fonction d'activation sigmoïde et un neurone de sortie linéaire.

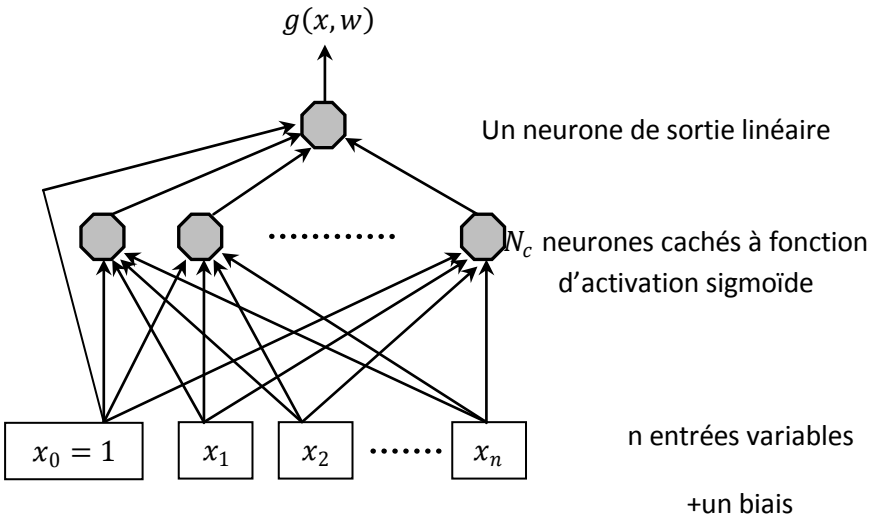


Figure 4.20.

Un réseau de neurones à n+1 entrées, une couche de N_c neurones cachés à fonction d'activation sigmoïde et un neurone de sortie linéaire.

La sortie de ce réseau a pour expression :

$$g(x,w) = \sum_{i=1}^{N_c} \left[w_{N_c+1,i} \tanh\left(\sum_{j=0}^{n} w_{ij}\, x_j \right) \right] + w_{N_c+1,0}$$

Sa sortie $g(x, w)$ est une fonction non linéaire du vecteur des entrées **x**, de composantes $1, x_1, x_2, \ldots, x_n$, et du vecteur des paramètres **w**, dont les composantes sont les $(n+1)N_c + N_c + 1$ paramètres du réseau.

Les neurones cachés sont numérotés de 1 à N_c et le neurone de sortie est numéroté $N_c + 1$. Par convention, le paramètre w_{ij} est relatif à la connexion allant du neurone *j* (ou de l'entrée j) vers le neurone i.

Très important

La sortie du réseau $g(x, w)$ est une fonction linéaire des poids de la dernière couche de connexions (qui relient les N_c neurones cachés au neurone de sortie, numéroté $N_c + 1$), et elle est une fonction non linéaire des paramètres de la première couche de connexions (qui relient les $\mathrm{n} + 1$ entrées du réseau aux N_c neurones cachés). Cette propriété a des conséquences importantes relativement à propriété concernant la parcimonie.

Propriété (1) :

Toute fonction bornée suffisamment régulière peut être approchée uniformément, avec une précision arbitraire, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire [HORNIK et al. 1989], [HORNIK et al. 1990], [HORNIK 1991].

Cette propriété garantie l'existence de la solution, mais elle ne fournit pas de méthode pour la trouver. Les questions qui restent posées sont les suivantes :

- Quel est le type de fonction sigmoïde le plus approprié ?
- Quel est l'algorithme avec lequel il faut entraîner le réseau ?
- Cet algorithme assure-t-il la convergence vers la solution ?
- En vue de considérations pratiques, existe-t-il un algorithme qui soit raisonnablement rapide ?
- Combien faut-il de neurones dans la couche cachée ?
- Quelle serait la qualité de la généralisation, c'est-à-dire la qualité de l'extrapolation en dehors de l'ensemble d'apprentissage ?

Nous allons prodiguer dans ce qui suit les réponses à ces questions avec ce qui leur convient comme arguments.

<u>En ce qui concerne la fonction d'activation</u>

Parmi les fonctions non linéaires, continues et bornées que sont les sigmoïdes, la tangente hyperbolique est préférée pour la raison suivante :

Un perceptron multicouche entraîné avec l'algorithme de rétro-propagation (que nous allons expliciter dans la suite) apprend plus vite (en termes de nombre d'itérations d'entraînement requis) quand la fonction d'activation sigmoïde du neurone est antisymétrique plutôt que quand elle est non symétrique. Cette condition n'est pas remplie par la fonction logistique standard.

La tangente hyperbolique est définie par

$$\varphi(v) = \mathrm{a}.\tanh(bv)$$

Où *a* et *b* sont des constantes.

L'algorithme avec lequel il faut entraîner le réseau

L'apprentissage du perceptron multicouche se fait en mode supervisé avec la règle delta ou (de correction d'erreur).

Ainsi, une fonction de coût se définit. Cette fonction de coût étant posée comme l'erreur quadratique moyenne.

Il s'agit donc de trouver quels sont les poids des connexions des neurones qui font aboutir à une minimisation de cette fonction de coût.

En cela, la connaissance de la valeur du gradient de cette fonction est nécessaire pour pouvoir rectifier les poids afin d'aller dans le sens de la descente de ce gradient.

Par conséquent, dans la procédure se dégagent deux étapes :

- L'évaluation du gradient
- L'ajustement des poids, suite à cette évaluation.

L'évaluation du gradient se fait

- soit par ce qui est appelé *l'algorithme de rétro-propagation* (et qui n'est pas en fait un algorithme d'apprentissage mais juste un ingrédient dans l'algorithme).
- Soit dans un sens direct.

Evaluation du gradient par rétro-propagation

Le neurone *i* calcule une grandeur y_j qui est une fonction non linéaire de son potentiel v_i ; le potentiel v_i est une somme pondérée des entrées x_j, la valeur de l'entrée x_j étant pondérée par un paramètre w_{ij} :

$$y_i = f\left(\sum_{j=1}^{n_i} w_{ij}\, x_j\right) = f(v_i)$$

Les n_i entrées du neurone *i* peuvent être soit les sorties d'autres neurones, soit les entrées du réseau.

La fonction de coût dont on cherche à évaluer le gradient est de la forme

$$J(w) = \sum_{k=1}^{N} \left(y_p^k - g\left(\boldsymbol{x^k}, \boldsymbol{w}\right)\right)^2 = \sum_{k=1}^{N} \boldsymbol{J^k}(\boldsymbol{w})$$

Pour évaluer son gradient, il suffit donc d'évaluer le gradient du coût partiel $\boldsymbol{J^k}(\boldsymbol{w})$ relatif à l'observation *k,* et de faire ensuite la somme sur tous les exemples.

L'algorithme de rétro-propagation consiste essentiellement en l'application répétée de la règle des dérivées composées. On remarque tout d'abord que la fonction de coût partielle ne dépend du

paramètre w_{ij} que par l'intermédiaire de la valeur de la sortie du neurone *i*, qui est elle-même fonction uniquement du potentiel du neurone *i* ; on peut donc écrire :

$$\left(\frac{\partial J^k}{\partial w_{ij}}\right)_k = \left(\frac{\partial J^k}{\partial v_i}\right)_k \left(\frac{\partial v_i}{\partial w_{ij}}\right)_k = \delta_i^k x_j^k$$

Où

$\left(\frac{\partial J^k}{\partial v_i}\right)_k$ désigne la valeur du gradient du coût partiel par rapport au potentiel du neurone *i* lorsque les entrées du réseau sont celles qui correspondent à l'exemple *k*.

$\left(\frac{\partial v_i}{\partial w_{ij}}\right)_k$ désigne la valeur de la dérivée partielle du potentiel du neurone *i* par rapport au paramètre w_{ij} lorsque les entrées du réseau sont celles qui correspondent à l'exemple *k*.

x_j^k est la valeur de l'entrée *j* du neurone *i* lorsque les entrées du réseau sont celles qui correspondent à l'exemple *k*.

Il reste donc à évaluer les quantités δ_i^k présentes dans le membre de droite de l'équation. Ces quantités seront calculées d'une manière récursive en menant les calculs depuis la sortie du réseau vers ses entrées.

Pour le neurone de sortie *i* ;

$$\delta_i^k = \left(\frac{\partial J^k}{\partial v_i}\right)_k = \frac{\partial}{\partial v_i}\left[\left(y_p^k - g(\boldsymbol{x}, \boldsymbol{w})\right)^2\right] = -2g(\boldsymbol{x^k}, \boldsymbol{w})\left(\frac{\partial g(\boldsymbol{x}, \boldsymbol{w})}{\partial v_i}\right)_k$$

Or, la sortie du réseau est la sortie y_i du neurone de sortie ; cette relation s'écrit donc :

$$\delta_i^k = -2g(\boldsymbol{x^k}, \boldsymbol{w}) f'(v_j^k)$$

Où $f'(v_j^k)$ désigne la dérivée de la fonction d'activation du neurone de sortie lorsque les entrées du réseau sont celles de l'exemple *k*. Si, comme c'est le cas, lorsque le réseau est utilisé en régression, le neurone de sortie est linéaire, l'expression se réduit à $\delta_i^k = -2g(\boldsymbol{x^k}, \boldsymbol{w})$

Pour un neurone caché *i* : la fonction de coût ne dépend du potentiel du neurone *i* que par l'intermédiaire des potentiels des neurones *m* qui reçoivent la valeur de la sortie du neurone *i*, c'est-à-dire de tous les neurones qui, dans le graphe des connexions du réseau, sont adjacents au neurone *i*, entre ce neurone et la sortie :

$$\delta_i^k = \left(\frac{\partial J^k}{\partial v_i}\right)_k = \sum_m \left(\frac{\partial J^k}{\partial v_m}\right)_k \left(\frac{\partial v_m}{\partial v_i}\right)_k = \sum_m \delta_m^k \left(\frac{\partial v_m}{\partial v_i}\right)_k$$

Or,

$$v_m^k = \sum_i w_{mi} x_i^k = \sum_i w_{mi} f(v_i^k)$$

D'où

$$\left(\frac{\partial v_m}{\partial v_i}\right)_k = w_{mi}\, f'\left(v_i^k\right)$$

Finalement, on obtient la relation :

$$\delta_i^k = \sum_m \delta_m^k\, w_{mi}\, f'\left(v_i^k\right) = f'\left(v_i^k\right) \sum_m \delta_m^k\, w_{mi}$$

Ainsi, les quantités δ_i^k peuvent être calculées récursivement, en parcourant le graphe des connexions à l'envers, depuis la sortie vers les entrées du réseau.

Une fois que les gradients des coûts partiels ont été calculés, il suffit d'en faire la somme pour obtenir le gradient de la fonction de coût totale.

Evaluation du gradient dans le sens direct

Dans cette méthode le calcul s'effectue dans le sens direct, en évaluant les gradients à partir des entrées vers la sortie.

- Pour un neurone *m* qui reçoit une information x_j^k directement de l'entrée j du réseau ou de la sortie du neurone *j* :

$$\left(\frac{\partial y_m}{\partial w_{mj}}\right)_k = \left(\frac{\partial y_m}{\partial v_m}\right)_k \left(\frac{\partial v_m}{\partial w_{mj}}\right)_k = f'\left(v_m^k\right) x_j^k$$

Où x_j^k est la valeur de l'entrée *j* du réseau pour l'exemple *k*.

Pour un neurone *m* qui reçoit une information x_j^k directement de l'entrée *j* du réseau ou de la sortie du neurone *j*, par l'intermédiaire d'autres neurones du réseau, situés entre les entrées et le neurone *m* :

$$\left(\frac{\partial y_m}{\partial w_{ij}}\right)_k = \left(\frac{\partial y_m}{\partial v_m}\right)_k \left(\frac{\partial v_m}{\partial w_{ij}}\right)_k = f'\left(v_m^k\right) \sum_l \left(\frac{\partial v_m}{\partial y_l}\right)_k \left(\frac{\partial y_l}{\partial w_{ij}}\right)_k = f'\left(v_m^k\right) \sum_l w_{ml} \left(\frac{\partial y_l}{\partial w_{ij}}\right)_k$$

Où l'indice *l* désigne tous les neurones qui sont adjacents au neurone *m* dans le graphe des connexions, entre le neurone *j* (ou l'entrée *j*) et le neurone *m*.

Ces deux relations permettent de calculer récursivement les dérivées de la sortie de chaque neurone par rapport aux paramètres qui ont une influence sur cette sortie, à partir des entrées du réseau jusqu'à la sortie de ce dernier.

Une fois toutes les dérivées calculées, on peut calculer le gradient de la fonction de coût partielle :

$$\left(\frac{\partial J^k}{\partial w_{ij}}\right)_k = \left(\frac{\partial}{\partial w_{ij}}\left[\left(y_p^k - g(\boldsymbol{x}, \boldsymbol{w})\right)^2\right]\right)_k = -2\left(y_p^k - g\left(\boldsymbol{x^k}, \boldsymbol{w}\right)\right)\left(\frac{\partial g(\boldsymbol{x}, \boldsymbol{w})}{\partial w_{ij}}\right)_k$$

Or, $g(\boldsymbol{x}, \boldsymbol{w})$ est la sortie d'un neurone du réseau, dont la dernière dérivée peut être calculée récursivement par le même procédé que toutes les autres. Une fois évalué le gradient du coût partiel pour chaque exemple, on fait la somme de ces gradients comme pour la rétro-propagation.

Remarque

La rétro-propagation nécessite l'évaluation d'un gradient par neurone, alors que le calcul direct requiert l'évaluation d'un gradient par connexion. Comme le nombre de connexions est à peu près proportionnel au carré du nombre de neurones, le nombre d'évaluations de gradient est plus important pour le calcul direct que pour la rétro-propagation.

Pour cette raison, il serait plus avantageux d'utiliser la rétro-propagation pour évaluer le gradient de la fonction de coût.

Modification des poids après l'évaluation du gradient de la fonction de coût

L'évaluation du gradient de la fonction de coût se fait à chaque itération du processus d'apprentissage. Après chaque évaluation, on effectue une modification des poids dans le but d'approcher le minimum de la fonction de coût.

Parmi les algorithmes de minimisation itérative figurent les méthodes du premier ordre et les méthodes de second ordre.

La méthode du gradient simple

Elle consiste à modifier les poids, à chaque itération, par la formule suivante :

$$w(i) = w(i-1) - \mu_i \nabla J\big(w(i-1)\big)$$

Avec $\mu_i > 0$.

La direction de descente est opposée à celle du gradient, c'est la direction suivant laquelle la fonction de coût diminue le plus rapidement. μ_i est appelé *pas du gradient* ou *pas d'apprentissage*.

Deux inconvénients de cette méthode

- Au voisinage d'un minimum de la fonction de coût, son gradient tend vers zéro, donc l'évolution du vecteur des coefficients devient très lente ; il en va de même si la fonction de coût présente des « plateaux » où sa pente est très faible. Il est impossible de savoir si une évolution très lente du gradient est due au fait que l'on est au voisinage d'un minimum, ou que l'on se trouve sur un plateau de la fonction de coût.
- Si la courbure de la surface de coût varie beaucoup, la direction du gradient peut être très différente de la direction qui mènerait vers le minimum ; c'est le cas si le minimum cherché se trouve dans une « vallée » longue et étroite (les courbes de niveau sont des ellipsoïdes allongées au voisinage du minimum).

Pour faire face à ces deux problèmes, on utilise des méthodes du second ordre qui, au lieu de modifier les coefficients uniquement en fonction du gradient de la fonction de coût, utilisent les dérivées secondes de cette dernière.

A souligner que les temps de convergence de la méthode du gradient simple sont supérieurs de plusieurs ordres de grandeur à ceux des méthodes du second ordre.

Les méthodes du gradient du second ordre

Toutes les méthodes du gradient de second ordre sont dérivées de la méthode de Newton dont voici le principe :

Le développement de Taylor d'une fonction $J(w)$ d'une seule variable *w*au voisinage d'un minimum w^* est donné par la relation :

$$J(w) = J(w^*) + \frac{1}{2}(w - w^*)^2 \left(\frac{d^2J}{dw^2}\right)_{w=w^*} + O(w^3)$$

Car le gradient de la fonction de coût est nul au minimum.

Une approximation du gradient de la fonction de coût au voisinage du minimum est obtenue en dérivant la relation précédente par rapport à *w* :

$$\frac{dJ}{dw} = (w - w^*)\left(\frac{d^2J}{dw^2}\right)_{w=w^*}$$

Par conséquent, lorsque la variable *w* est au voisinage de w^*, on pourrait atteindre ce minimum en une seule itération si l'on connaissait la dérivée seconde de la fonction à son minimum : il suffit pour cela de modifier la variable *w* de la quantité

$$\Delta w = \frac{{dJ}/{dw}}{\left({d^2J}/{dw^2}\right)_{w=w^*}}$$

Le même raisonnement s'applique à une fonction de plusieurs variables, la dérivée seconde étant remplacée par la matrice hessienne $H(w)$de la fonction à optimiser, de terme général$\frac{\partial^2 J}{\partial w_i \partial w_j}$.

Pour atteindre le minimum de la fonction de coût en une itération, il suffirait d'appliquer au vecteur des poids la modification suivante (sous réserve que la matrice hessienne soit inversible) :

$$\Delta w = -H(w^*)^{-1}\nabla J(w)$$

Cette formule n'est pas applicable en pratique du fait que w^* n'est pas connu. Néanmoins, il existe plusieurs techniques qui mettent en œuvre une approximation itérative de la matrice hessienne (ou de son inverse). Parmi ces techniques, celle qui est indiquée lorsque le réseau est petit et lorsque la fonction de coût est quadratique, est celle connue sous le nom de l'algorithme de Levenberg-Marquardt.

L'algorithme de Levenberg-Marquardt

L'algorithme de Levenberg-Marquardt consiste à modifier les paramètres par la formule :

$$w(i) = w(i-1) - \left[H\big(w(i-1)\big) + \mu_i I\right]^{-1} \nabla J\big(w(i-1)\big)$$

Pour de petites valeurs du pas μ_i, la méthode de Levenberg-Marquardt s'approche de celle de Newton. Inversement, pour de grandes valeurs de μ_i, l'algorithme de Levenberg-Marquardt est équivalent à l'application de la règle du gradient simple avec un pas de $1/\mu_i$. L'application de cet algorithme nécessite l'inversion de la matrice $\left[H\big(w(i-1)\big) + \mu_i I\right]$. L'expression exacte de la matrice hessienne de la fonction de coût totale $J(w)$ est :

$$H\big(w(i)\big) = \sum_{k=1}^{N} \left(\frac{\partial e^k}{\partial w(i)}\right)\left(\frac{\partial e^k}{\partial w(i)}\right)^T + \sum_{k=1}^{N} \frac{\partial^2 e^k}{\partial w(i)\partial w(i)^T} e^k$$

Avec $e^k = y_p^k - y^k$

Le second terme de cette expression étant proportionnel à l'erreur, on peut le négliger en première approximation, ce qui fournit une approximation approchée :

$$\widetilde{H}\big(w(i)\big) = \sum_{k=1}^{N} \left(\frac{\partial e^k}{\partial w(i)}\right)\left(\frac{\partial e^k}{\partial w(i)}\right)^T = \sum_{k=1}^{N} \left(\frac{\partial y^k}{\partial w(i)}\right)\left(\frac{\partial y^k}{\partial w(i)}\right)^T$$

Dans le cas d'un modèle linéaire par rapport aux paramètres, *y* est une fonction linéaire de ***w***, donc le second terme de l'expression de $\widetilde{H}$ est nul : l'approximation devient exacte.

Sélection de modèle

La propriété (1) évoquée précédemment garantie l'existence d'une solution par le moyen de l'architecture qui y est décrite. Néanmoins, elle ne fournit pas une procédure qui permet de déterminer exactement cette architecture.

Il s'agit de sélectionner le meilleur modèle.

Dans notre cas, la question devient : combien la couche cachée doit-elle comporter de neurones ?

Le nombre de neurones cachés représente la complexité du modèle.

Chaque neurone ajoute des paramètres au modèle. Plus il y a de paramètres, plus le modèle devient flexible. C'est en ce point que surgit le dilemme biais-variance.

Un modèle très complexe, possédant un très grand nombre de paramètres ajustables peut avoir un biais très faible, c'est-à-dire s'ajuster aux données quel que soit le bruit présent dans celles-ci, mais il risque d'avoir une variance très grande, c'est-à-dire être très dépendant de la réalisation particulière du bruit présent dans les données d'apprentissage.

A l'inverse, un modèle très simple, possédant peu de paramètres ajustables, peut être très peu dépendant du bruit présent dans les données, mais s'avérer incapable d'être proche de la régression.

En termes appropriés aux réseaux de neurones, nous pouvons dire qu'il faut trouver un modèle qui réalise le meilleur compromis entre les capacités d'apprentissage et les capacités de généralisation : si le réseau apprend « trop bien », il s'ajuste au bruit, et donc a de mauvaises performances de généralisation

Il ya exigence de concilier entre ces deux contraintes.

Le biais et la variance, comme la fonction de coût théorique, n'étant pas calculables, le difficile problème de la sélection de modèle consiste donc à essayer de résoudre la contradiction entre biais et variance sans que l'on puisse calculer ces deux grandeurs.

L'approche consacrée est de recourir à une procédure en deux étapes :

- Un processus constructif, par augmentation progressive de la complexité du modèle.
- Pour une famille de même complexité, on effectue plusieurs apprentissages, utilisant la totalité des exemples disponibles, avec des initialisations différentes des paramètres. Le choix se fixe sur le modèle qui réalise la plus petite valeur de la fonction de coût en même temps que des valeurs voisines du score de généralisation et celui de l'entraînement.

CHAPITRE V

ESTIMATION DU DÉLAI POST-MORTEM PAR LE MOYEN DES RÉSEAUX DE NEURONES ARTIFICIELS

Dans ce chapitre nous proposons une nouvelle méthode pour l'estimation du délai post-mortem. Nous utilisons en cela des réseaux de neurones artificiels à propagation avant et à apprentissage supervisé. Nous avons mené une étude comparative sur un échantillon de 257 individus pour souligner, par rapport à la formule de Henssge, l'avantage considérable qu'apporte cette nouvelle technique dans la précision des estimations.

1. Introduction

Un problème important qui se pose en médecine légale est celui de l'estimation du délai post mortem. Des méthodes ont été successivement développées dans le but d'aboutir à des estimations acceptables. Celle qui est la plus en vue est la méthode thermométrique, qui continue de constituer l'un des principaux outils pour l'évaluation de l'intervalle post-mortem. Son intérêt repose sur le fait qu'elle soit construite sur la base de mesures quantitatives.

Une série d'hypothèses successives, surclassant l'une l'autre, ont conduit à son évolution :

- le refroidissement cadavérique est proportionnel au temps.
- le flux thermique est proportionnel à la différence de température entre le corps et l'air ambiant.

Malheureusement, les observations réelles montrent que les modèles qui en ont résulté, linéaire pour la première hypothèse et exponentiel pour la seconde, demeurent très imprécis.

Le modèle le plus récent et le plus utilisé actuellement est bâti sur la formule proposée par C.Hennsge [1]. :

$$\frac{T_{corps} - T_{ambiant}}{37{,}2 - T_{ambiant}} = 1{,}25.e^{-kt} - 0{,}25.e^{-5kt}$$

où k est un paramètre dépendant de la masse M (en kg) de l'individu :

$$\mathrm{k} = \frac{1{,}2815}{\mathrm{M}^{0{,}625}} - 0{,}0284$$

Cette formule tient compte de trois facteurs : la température ambiante, la température du corps et la masse corporelle.

Le temps (t) marquant le délai entre la survenue de la mort et le moment des mesures des températures et du poids, est déduit de cette formule sous les conditions suivantes :

- Air sec sans mouvements
- Corps totalement nu depuis le moment de la mort.

Mais lorsque ces conditions ne sont pas respectées, un coefficient correctif est appliqué pour tenir compte des diverses variantes.

La méthode de Henssge, bien que meilleure que celles qui l'ont précédé, ne résout pas tout à fait le problème, l'intervalle d'estimation (plus de trois heures) reste trop large pour répondre correctement à l'exigence des considérations pratiques.

Le problème est que la loi sous jacente au phénomène doit être beaucoup trop complexe pour être saisie par cette formule. Le modèle est manifestement hautement non linéaire, et il n'existe pas encore de connaissances théoriques suffisantes pour déterminer sa structure.

Cependant, le problème de l'estimation du délai post-mortem est un problème typique qui entre parfaitement dans la catégorie pour laquelle les réseaux de neurones constituent d'excellents outils de modélisation non linéaire par apprentissage.

La fonction qui peut lier dans la réalité le délai post-mortem aux différents paramètres qui sont le poids du corps, la température rectale et la température ambiante pourrait être assez complexe par rapport à celle proposée par Henssge. Cependant, il est tout à fait concevable qu'elle ne peut être que bornée et suffisamment régulière par rapport aux différentes variables (ne présentant pas de discontinuités ou des points anguleux). C'est ce qui la rend candidate à être approchée autant que l'on veut, moyennant une quantité de données suffisante et un nombre de neurones adéquat. Ceci est permis par la propriété (1) (énoncée au chapitre précédent) dont jouissent les réseaux de neurones non bouclés à apprentissage supervisé :

''Toute fonction bornée suffisamment régulière peut être approchée uniformément, avec une précision arbitraire, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire

La formule de Henssge est utilisée pour estimer le délai post-mortem (DPM) à partir des valeurs du poids (P), de la température rectale(T_r°) et de la température ambiante (T_a°).

Nous avons utilisé des réseaux de neurones non bouclés à apprentissage supervisé pour effectuer cette même tâche d'estimation. Comme pour la formule de Henssge, (P), (T_r°) et (T_a°) constituent les entrées et (DPM) constitue la sortie.

L'objet de l'étude est de comparer la performance de la formule de Henssge et celle des réseaux de neurones, en termes d'écart de l'estimation du DPM par rapport à sa vraie valeur.

2. Etude expérimentale

2.1. Description du réseau

Le réseau utilisé est un réseau non bouclé à propagation avant. Il est doté de :

- une couche cachée de neurones ayant tous la tangente hyperbolique comme fonction d'activation.
- Un neurone de sortie dont la fonction d'activation est linéaire.

Le mode d'apprentissage est l'apprentissage supervisé où les cibles sont fournies au réseau en même temps que les entrées. Dans cette phase d'entraînement, le réseau doit comparer ses propres réponses aux valeurs cibles et ajuster, en conséquence et d'une manière itérative, ses paramètres (les Poids synaptiques) pour s'en rapprocher.

L'algorithme d'apprentissage est l'algorithme de Levenberg-Marquart.

Les observations sont réparties aléatoirement en trois ensembles :

- Le premier ensemble, constitué de 60% des observations est utilisé pour l'entraînement
- Le deuxième ensemble, constitué de 20% des observations est utilisé pour la validation (la capacité de généralisation). L'entraînement s'arrête automatiquement lorsqu'il y a risque de surajustement.
- Le troisième ensemble, constitué de 20% des observations est utilisé comme ensemble complètement indépendant pour mesurer la capacité de généralisation.

Deux critères (équivalents) sont utilisés pour mesurer la performance de chacune des deux méthodes :

- L'erreur moyenne (EM): qui est la moyenne des écarts des estimations par rapport aux vraies valeurs.
- L'erreur quadratique moyenne (EQM): qui est la moyenne des carrés de ces mêmes écarts.

2.2. Description des données

Les mesures sur l'ensemble des individus ont été faîtes dans les mêmes conditions :

- Air sec sans mouvements
- Corps totalement nu depuis le moment de la mort.

Le poids (P) des individus varie entre 3.5 Kg à 102 Kg

La température ambiante varie entre 4.5°C et 18°C

La température des corps varie entre 17°C et 37°C

Ces intervalles constituent, respectivement pour chacune des variables, les domaines d'apprentissage des réseaux. (Il est attendu que lorsqu'on s'éloigne du domaine d'apprentissage, la précision de l'approximation se dégrade). Des domaines d'apprentissage plus étendus exigent des données spécifiques récoltées à cet effet ; ce qui fera l'objet d'une étude ultérieure.

2.3. Les résultats

Le délai post-mortem, intervalle de temps entre le moment exact de la mort et la prise de la température rectale, a été mesuré puis converti en valeurs décimales.

Les observations ont été classées en fonction du délai post-mortem réel par ordre croissant, de la plus petite (1.33) à la plus grande (18.33), et ont été réparties en trois catégories :

- Petites valeurs : de 1.33 à 3.94 (70 observations)
- Valeurs moyennes : de 4.00 à 6.92 (107 observations)
- Grandes valeurs : de 7.00 à 18.33 (80 observations)

Nous avons entraîné et spécialisé quatre types de réseaux : Des réseaux sur chacune des trois catégories et des réseaux sur l'ensemble des observations.

L'étude a révélé que le nombre optimal de neurones dans la couche cachée est 10. C'est ce nombre qui a assuré, en fonction de la quantité de données dont nous disposons, le meilleur compromis entre la précision de l'estimation et une bonne généralisation.

Tous les réseaux dont les résultats sont présentés ci-après ont 10 neurones dans la couche cachée et un neurone de sortie.

Dans chaque catégorie, nous avons retiré par une procédure aléatoire quelques observations pour les besoins de tests supplémentaires complètement indépendants de la procédure de calcul.

L'évaluation est faite sur deux critères :

- Erreur moyenne et erreur quadratique moyenne sur l'échantillon d'étude principal.
- Erreur moyenne et erreur quadratique moyenne sur l'échantillon du test supplémentaire.

1. PETITES VALEURS (de 1.33 à 3.94):

Première étude : L'échantillon pour le test est choisi aléatoirement mais parmi ceux plus favorable que l'échantillon principal.

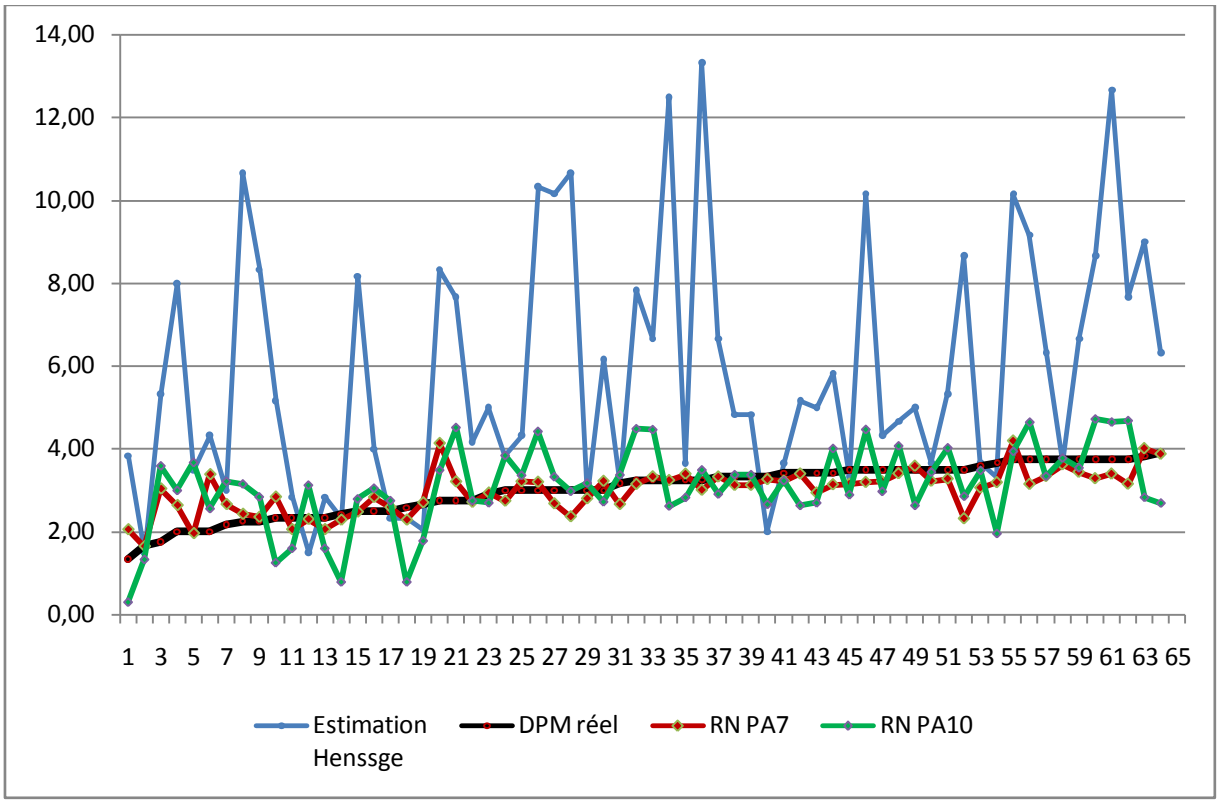
a) Etude principal

Echantillon : 64 observations.

Performances :

- Formule de Henssge : E M = 2.94 EQM = 15.92
- Réseaux de neurones :

Réseaux	PA1	PA2	PA3	PA4	PA5	PA6	PA7	PA8	PA9	PA10
EM	0.60	0.37	0.39	0.35	0.35	0.39	0.32	0.37	0.63	0.70
EQM	0.55	0.27	0.29	0.26	0.47	0.28	0.20	0.23	0.60	0.72



b) Test supplémentaire

Echantillon : 6 observations.

Performances :

- Formule de Henssge : E M = 1.76 EQM = 4.32
- Réseaux de neurones :

Réseaux	PA1	PA2	PA3	PA4	PA5	PA6	PA7	PA8	PA9	PA10
EM	0.40	0.60	0.50	0.62	0.60	0.64	0.39	0.40	0.44	0.96
EQM	0.34	0.55	0.31	0.49	0.53	0.52	0.20	0.20	0.23	1.02

Deuxième étude

L'échantillon pour le test est choisi aléatoirement mais parmi ceux moins favorable que l'échantillon principal.

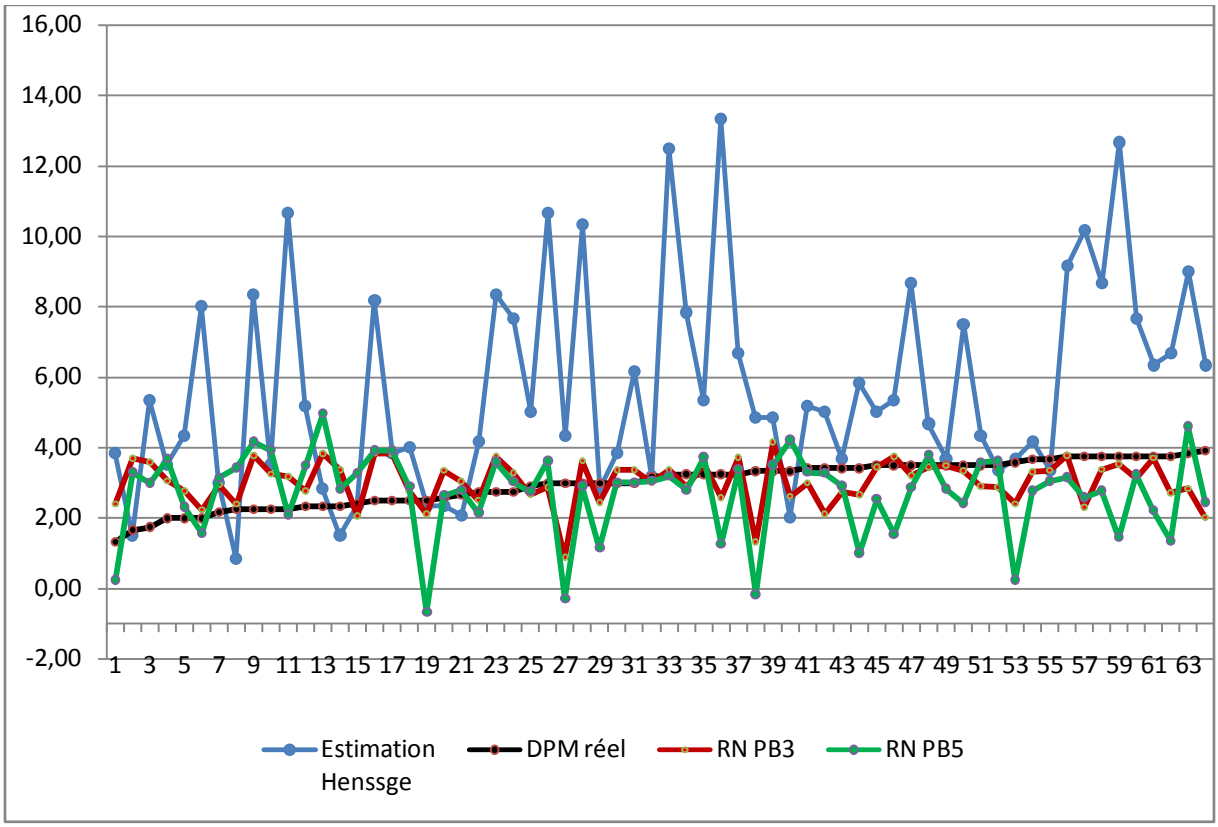
a) Etude principal

Echantillon : 64 observations.

Performances :

- Formule de Henssge : E M = 2.82 EQM = 14.65
- Réseaux de neurones :

Réseaux	PB1	PB2	PB3	PB4	PB5	PB6	PB7	PB8	PB9	PB10
EM	0.46	0.53	0.22	0.40	0.79	0.43	0.44	0.41	0.38	0.47
EQM	0.57	0.45	0.15	0.27	1.00	0.33	0.31	0.31	0.36	0.37



b) Test supplémentaire :

Echantillon : 6 observations.

Performances :

- Formule de Henssge : E M = 3.03 EQM = 17.89
- Réseaux de neurones :

Réseaux	PA1	PA2	PA3	PA4	PA5	PA6	PA7	PA8	PA9	PA10
EM	0.50	0.45	1.55	0.66	1.03	0.56	0.27	0.84	0.97	0.33
EQM	0.33	0.32	8.33	0.95	2.04	1.24	0.08	1.22	2.59	0.16

2. VALEURS MOYENNES (de 4.00 à 6.92) :

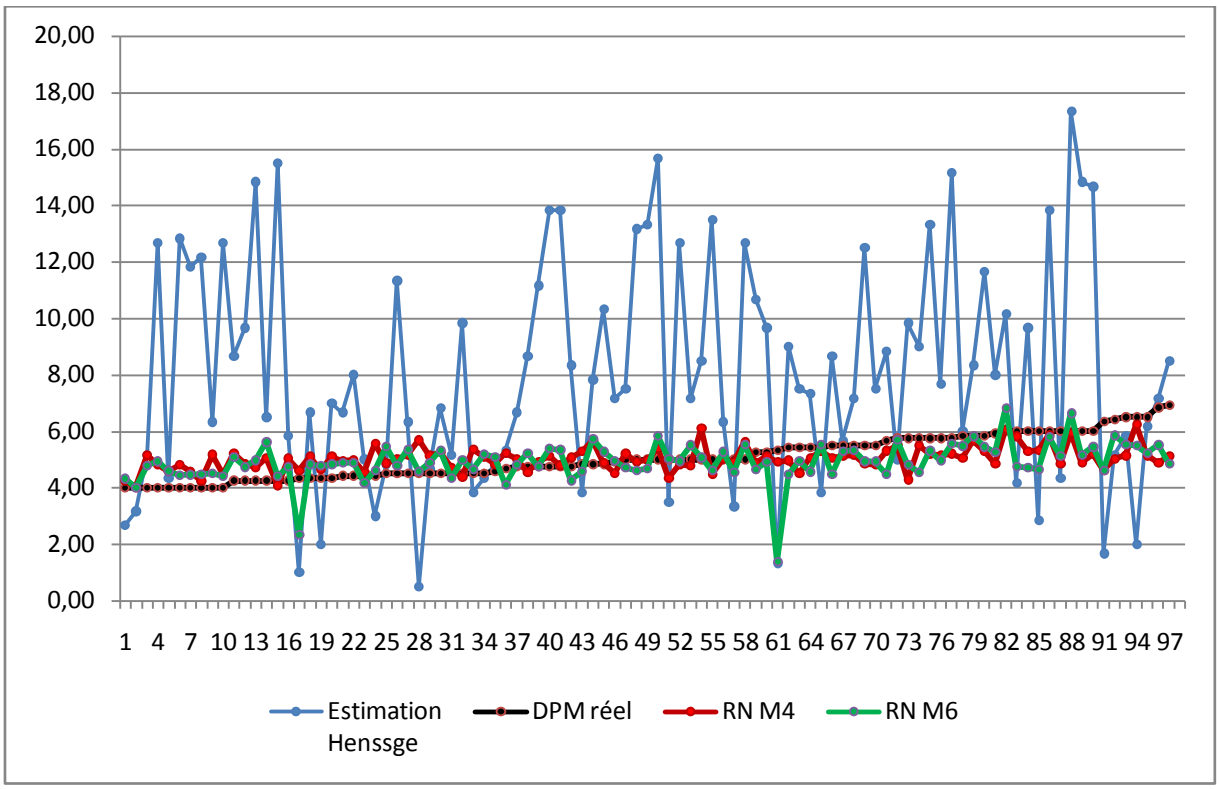
a) Etude principal

Echantillon : 97 observations.

Performances :

- Formule de Henssge : E M = 3.89 EQM = 24.62
- Réseaux de neurones :

Réseaux	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
EM	0.61	0.51	0.59	0.57	0.62	0.63	0.61	0.65	0.59	0.58
EQM	0.56	0.54	0.60	0.50	0.54	0.68	0.62	0.66	0.61	0.56



b) Test supplémentaire

Echantillon : 10 observations.

Performances :

- Formule de Henssge : E M = 2.86 EQM = 14.14
- Réseaux de neurones :

Réseaux	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
EM	0.75	1.02	0.86	0.78	0.79	0.71	0.99	1.10	0.75	0.82
EQM	0.69	1.22	0.92	0.75	0.76	0.88	1.06	1.36	0.90	0.87

3. GRANDES VALEURS (de 7.00 à 18.33) :

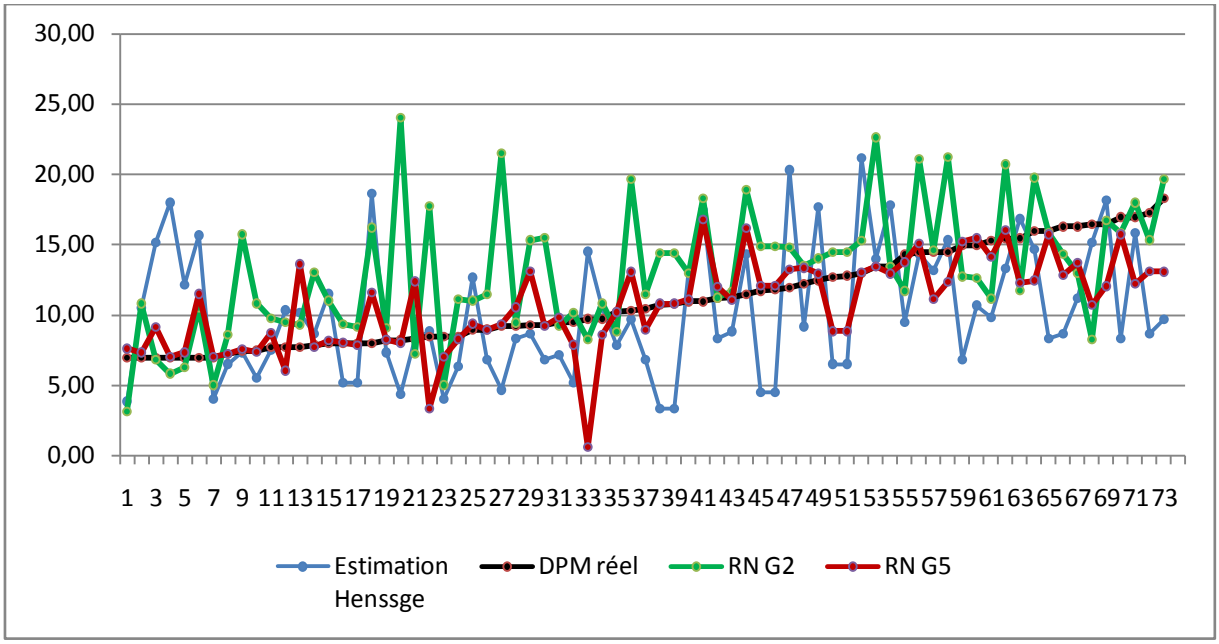
a) Etude principal

Echantillon : 73 observations.

Performances :

- Formule de Henssge : E M = 3.92 EQM = 23.54
- Réseaux de neurones :

Réseaux	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
EM	2.78	3.63	3.27	3.20	1.73	1.96	3.95	2.97	2.37	2.17
EQM	12.58	25.90	20.18	19.03	6.97	8.22	24.78	12.85	10.04	9.98



b) Test supplémentaire

Echantillon : 7 observations.

Performances :

- Formule de Henssge : E M = 2.42 EQM = 9.26
- Réseaux de neurones :

Réseaux	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
EM	2.63	5.05	3.56	2.04	2.85	4.70	2.02	2.42	2.79	2.98
EQM	9.04	41.81	16.08	6.28	14.12	29.55	10.41	10.57	12.54	10.51

4. TOUTES LES VALEURS

Echantillon principal : 237 observations

Echantillon du test supplémentaire : 20 observations

Nous nous limitons ici à reporter les résultats de la comparaison de l'erreur quadratique moyenne (EQM) produite par la formule de Henssge sur l'échantillon du test supplémentaire et celles de 40 réseaux sur le même échantillon.

Performances :

- ErreursQuadratique Moyenne de la formule de Henssge : EQM = 16.33
- Erreurs Quadratiques Moyennes des réseaux de neurones :

Réseau T1	Réseau T2	Réseau T3	Réseau T4	Réseau T5	Réseau T6	Réseau T7	Réseau T8	Réseau T9	Réseau T10
5.16	4.64	8.66	4.67	4.83	4.68	6.95	6.71	5.38	4.38
Réseau T11	Réseau T12	Réseau T13	Réseau T14	Réseau T15	Réseau T16	Réseau T17	Réseau T18	Réseau T19	Réseau T20
6.17	8.4	6.52	4.72	7.75	12.45	7.26	4.25	5.56	6.12
Réseau T21	Réseau T22	Réseau T23	Réseau T24	Réseau T25	Réseau T26	Réseau T27	Réseau T28	Réseau T29	Réseau T30
4.94	5.29	3.98	7.91	4.06	5.44	6.83	5.36	5.00	4.37
Réseau T31	Réseau T32	Réseau T33	Réseau T34	Réseau T35	Réseau T36	Réseau T37	Réseau T38	Réseau T39	Réseau T40
5.03	4.32	3.63	6.50	5.38	4.58	5.64	5.08	5.86	4.41

En entraînant les réseaux sur toutes les valeurs (237) nous obtenons des résultats de qualité intermédiaire entre celle où les réseaux ont de bonnes performances (ceux liés aux petites valeurs et aux valeurs moyennes) et celle où les réseaux ont de mauvaises performances (ceux liés aux grandes valeurs).

3. Conclusion

Cette étude prouve que, pour déjà des échantillons de cette dimension réduite, les estimations obtenues par les réseaux de neurones sont de loin plus satisfaisantes que celles obtenues par la formule de Henssge. Elles le sont particulièrement pour le premier palier de refroidissement (les 7 premières heures après la mort). C'est ce palier qui posait justement problème pour les méthodes traditionnelles du fait de la lenteur de la variation thermique.

Avec les données dont on dispose, les réseaux de neurones ne sont contre performants que pour les grandes valeurs (supérieurs à 7 heures).

Le remède à cela est d'augmenter le nombre d'observations entrant dans ce domaine d'apprentissage ou, à défaut, nous satisfaire des estimations faites par les réseaux entraînés sur toutes les observations. Néanmoins, les estimations données par ces derniers réseaux, bien que plus précises que celles obtenues par la formule de Henssge, ne sont pas de la même qualité que celles des réseaux entraînés pour les deux premières catégories spécifiées.

4. Extension de l'étude

Autant le nombre d'observations est grand autant les réseaux de neurones améliorent la précision des estimations et la fiabilité de la généralisation. Avantageusement par rapport aux méthodes traditionnelles qui produisent des formules figées et dont la performance reste constante, les réseaux de neurones ont la possibilité de s'améliorer continuellement.

Du fait de la mise en œuvre pratique et tout à fait aisée, l'entraînement peut être repris chaque fois que nous disposons de données supplémentaires.

Cette méthode est valable pour la thermométrie, mais elle l'est également tant que les variables prédictives sont quantitatives et tant que le modèle sous jacent est non linéaire. Les machines adaptatives offrent de très larges possibilités. Il semble, en effet, que des travaux récents ont eu recours à des outils adaptés aux modèles non linéaires (Support Vector Machines) afin de prédire le délai post mortem à partir des concentrations de certains éléments chimiques dans l'humeur vitrée [2].

La méthode que nous avons proposé, ne nécessitant pas un modèle de connaissance a priori, ouvre des champs d'investigations dans plusieurs directions et qui peuvent même sortir du cadre strict de la thermométrie. Nous pouvons envisager de :

- Renforcer les résultats sur les mêmes domaines d'apprentissage.
- Etendre ces domaines d'apprentissage, particulièrement celui de la température ambiante (au-delà de 18°C).
- Inclure des variables impliquant, d'une manière directe ou indirecte, la surface du corps et son volume.
- Inclure des variables quantifiables autres que la température (concentrations d'éléments chimiques).
- Vérifier l'influence de l'âge et du sexe.
- Quantifier et intégrer les caractéristiques du milieu.
- Vérifier l'influence de la forme médico-légale de la mort (violente et non violente).

Une chose qui reste très importante est que, en l'absence de connaissances théoriques sur le vrai modèle thermométrique, tout reste ouvert. Les nouvelles techniques, qui sont les machines apprenantes viennent à point nommé pour nous libérer de ce préalable. Cela ne sert pas uniquement pour l'estimation, mais aussi et surtout pour la **prospection**.

En effet, parmi toutes les variables quantitatives, quelles sont celles qui ont un effet sur le délai post mortem ?

Doit-on analyser l'effet isolé d'un facteur ou l'effet combiné de plusieurs facteurs?

Le champ est aujourd'hui ouvert à tout ce qui quantifiable, et l'étude peut sortir du cadre strict de la thermométrie pour englober des variables d'une autre nature; certains éléments chimiques par exemple, dans l'œil ou ailleurs.

D'autre part, dans la mesure où il y a plusieurs points de prise de la température autre que rectale (tête, abdomen, ...). La question qui surgit alors est la suivante:

Nous pouvons admettre que la température observe nécessairement une décroissance en tous ces points, mais est-ce de la même manière? est-ce d'une manière corrélée?

La possession de données, en quantité suffisante, sur des points multiples peut nous permettre d'avancer des hypothèses. Nous pourrons confirmer certaines et infirmer d'autres.

Ensuite, si nous avons utilisé la température rectale dans notre travail c’est pour rester dans les mêmes conditions de Henssge, puisque l'étude est comparative. Mais qui nous prouve qu'un autre point de mesure n'est pas plus pertinent?

REFERENCES

Théorie de l'information

1.FISHER, R. A., 1922, "On the Mathematical Foundations of Theoretical Statistics", Philosophical Transactions of the Royal Society, 222A, 1922, pp. 309-368

2. NYQUIST, H., 1924, "Certain Factors Affecting Telegraph Speed", Bell System Technical Journal, 3, April 1924, pp. 324-346

3. LEWIS, G.N., 1930, "The Symmetry of time in Physics", Science, 71, 1930, pp. 569-576

4. FISHER, R.A., 1925a, Statistical Methods for Research Workers, London, 1925 (14ème éd., Hafner, New York, 1973)

5. FISHER, R.A. 1925b, "TheTheory of Statistical Estimation", Proceedings of the Cambridge Philosophical Society, 22, 1925, pp. 700-725

6. FISHER 1934a : R. A. FISHER, "The amount of information supplied by records of families as a fonction of the linkage in the population sampled", Annals of Eugenics, 6, 1934, pp. 66-7

7. FISHER, R.A. 1934b, "Probability Likelihood and Quantity of Information in the logic of Uncertain Inference", Proceedings of the Royal Society A, 146, 1934, pp.1-8

8. FISHER,R.A., 1934c, "Indeterminism and Natural Selection", Philosophy of Science, 1, 1934, pp. 99-117

9. FISHER,R.A., 1935a, The Design of Experiments, 1935, 9ème édition, New-York, 1971

10. FISHER, R.A., 1935b, "The logic of inductive inference", Journal of the Royal Statistical Society, 98, 1935, 39-54 et discussion pp. 55-82

11. FISHER, R.A. 1956, Statistical Methods and Scientific Inference, Adelaïde, 1956

12. DOOB, J.L., 1936, "Statistical Estimation", Transactions of the American Mathematical Society, 39, 1936, pp.410-21

13. DOOB, J.L., 1941, "Probability as Measure", Annals of Mathematical Statistics, 12, 1941, pp.206-14

14. BHATTACHARYYA 1946, 1948, "On some analogues of the amount of information and their use in statistical estimation", Sankhia, 8, 1946, pp. 1-14; 8, 1947 p. 201-218; 8, 1948, pp. 315-328

15. HARTLEY, R.V.L., 1928, "Transmission of Information", Bell System Technical Journal, 7, 1928, pp 535-563

16. SHANNON, C.E. 1948, "A Mathematical Theory of Communication", Bell System Technical Journal, 27, 1948, pp. 379-423 et 623-656

17. KHINCHIN, A.I. 1953, 1956, Mathematical Foundations of Information Theory, Dover Pub. Inc. , New York, 1957 (Ed. russe : 1953 et 1956, traduit en RDA en 1957)

18. KOLMOGOROV, A.N., 1965, "Three Approaches of Quantitative Definition of Information", Problems of Information Transmission, 1, 1965, pp. 3-11

19. KOLMOGOROV, A.N., 1968, "Logical Basis for Information Theory and Probability Theory", IEEE Trans. on Information Theory, 14, 1968, pp. 662-664

19 (bis). KOLMOGOROV, A. ,1983, "Combinatorial foundations of information theory and the calculus of probabilities", Russian mathematical surveys, 38, 1983, pp. 29-40

20. CHAITIN, G., 1977, "Algorithmic Information Theory", IBM Journal of Research and Development, 21, 1977, pp. 350-359

psychiatrique, 20, 1949, Fasc IV, pp. 585-607

21(bis). SCHÜTZENBERGER: M.P., 1951, "Sur les rapports entre la quantité d'information au sens de Fisher et au sens de Wiener", Comptes rendus de l'Académie des Sciences, 232, 1951, pp. 925-927

22. SCHÜTZENBERGER, M.P., 1953, Contributions aux applications statistiques de la théorie de l'information, Publications de l'Institut de Statistique de l'Université de Paris, 3, 1953, pp. 3-117

22 (bis) SCHÜTZENBERGER, M.P., 1956b, "On some Measures of Information used in Statistics", in Information Theory, Papers read at a Symposium held at the Royal Institution, 12-16 Sept. 1955, Ed. C. Cherry, Butterworths Scientific Publications,

22 (ter). SCHÜTZENBERGER, M.P.,1983, "Théorie de l'information", in Information et Connaissance, `Recherches Interdisciplinaires du Collège de France', sous la Direction de A. Lichnerowicz et F. Perroux, Maloine, Paris, 1983

23. MANDELBROT, B., 1952b, "Sur la notion générale d'information et la durée intrinsèque d'une stratégie", Comptes rendus de l'Académie des Sciences, 234, 1952, 1345-6

24. KULLBACK, S., et LEIBLER, R.A., 1951, "On Information and Sufficiency", Annals of Mathematical Statistics, 22, 1951, pp. 79-86

25. KULLBACK, S., 1959, 1968, Information Theory and Statistics, John Wiley & Sons, New York, 1959 (2ème édition augmentée : Dover, New York, 1968)

26. JAYNES, E.T., 1957, "Information Theory and Statistical Mechanics" Physical Review, 106, 1957, pp. 620-630 et 108, 1957, pp. 171-190

26 (bis). JAYNES, E.T. , 1979 , "Information Theory and Statistical Mechanics" in The Maximum Entropy Formalism, R.D. LEVINE et M. TRIBUS, eds, Cambridge, M.I.T. Press, 1979.

27. FEINSTEIN, A., 1958, Foundations of Information theory, Mc Graw Hill, New-York, 1958

28. McMILLAN, B., and SLEPIAN, D., 1962, "Information Theory", Proceedings of the IRE, 50, 1962, pp. 1151-1157

29. WOODWARD, P.M., 1953, "Information Theory", British Journal of Applied Physics, 4, 1953, pp. 129-33

30. KOLMOGOROV, A ., 1956, "On the Shannon Theory of Information Transmission in the Case of Continuous Signals", IEEE Transactions on Information Theory, 2, 1956, pp. 47-60

31. SOLOMONOFF, R.J., 1964, "A formal Theory of Inductive Inference", Information and Control, 7, 1964, pp. 1-22 et 224-254

32. .JAYNES, E.T. , 1979 , "Information Theory and Statistical Mechanics" in The Maximum Entropy Formalism, R.D. LEVINE et M. TRIBUS, eds, Cambridge, M.I.T. Press, 1979

33. SCHÜTZENBERGER, M.P., 1951, "Sur les rapports entre la quantité d'information au sens de Fisher et au sens de Wiener", Comptes rendus de l'Académie des Sciences, 232, 1951, pp. 925-927

34. SCHÜTZENBERGER, M.P., 1953 , Contributions aux applications statistiques de la théorie de l'information, Publications de l'Institut de Statistique de l'Université de Paris, 3, 1953, pp. 3-117

35. SCHÜTZENBERGER, M.P., 1956b, "On some Measures of Information used in Statistics", in Information Theory, Papers read at a Symposium held at the Royal Institution, 12-16 Sept. 1955, Ed. C. Cherry, Butterworths Scientific Publications, London, Academic Press, New York, 1956, pp. 18-25

36. SCHÜTZENBERGER, M.P.,1983, "Théorie de l'information", in Information et Connaissance, 'Recherches Interdisciplinaires du Collège de France', sous la Direction de A. Lichnerowicz et F. Perroux, Maloine, Paris, 1983

37. SEGAL, J., 1998 , Théorie de l'information : sciences, techniques et société - de la seconde guerre mondiale à l'aube du XXIe siècle, Thèse de Doctorat, Faculté d'Histoire de l'Université Lyon II

38. SHANNON, C.E., 1948, "A Mathematical Theory of Communication", Bell System Technical Journal, 27, 1948, pp. 379-423 et 623-656

38(bis). SHANNON, C.E., 1964, "Information Theory", 12th Ed. of the Encyclopaedia Britannica, William Benton Publishers, Chicago, 1964

39. GUIACU, S., THEODORESCU, R., 1971, Incertitude et information,, Les presses de l'Université LAVAL, Québec.

40. VOLLE, M., 1993, Analyses des données, 3ème édition, Ed.ECONOMICA, PARIS

41. WEHENKEL, L., 2003, Théorie de l'information et du codage, notes de cours, Université de Liège, Faculté des sciences appliquées

Réseaux de neurones

42. McCULLOCH, W., et PITTS, W., 1943, "*A logical Calculus of the Ideas Immanent in Nervous Activity*", Bulletin of Mathematical Biophysics, 5, 1943, pp. 115-133

43. Hebb, D.O., 1949, *The Organization of Behavior* ; A Neuropsychological Theory, New York ; Wiley

44. Rochester, N., J.H. Holland, L.H.Haibt, and W.L. Duda, 1956 ; ''*Tests on a cell assembly theory of the action of the brain using a large digital computer*'', IRE Transactions on Information Theory, volIT-2, pp. 80-93

45. Uttley, A.M., 1956, '' *A theory of the mechanism of learning based on the computation of conditional probabilities*'', Proceedings of the First International Conference on Cybernetics, Namur, Gauthier-Villars, Paris.

46. Uttley, A.M., 1979, *Information Transmission in the Nervous System*, London ; Academic Press.

47. Ashby, W.R., 1952, Design for Brain, New York ; Wiley.

48. Minsky, M.L., 1954, ''*Theory of neural-analog reinforcement systems and its application to the brain-model problem*'', Ph.D. thesis, Princeton University, Princeton, NJ.

49. Minsky, M.L., 1961, ''*Steps towards artificial intelligence*'', Proceedings of the Institue of Radio Engineers, vol. 49, pp.8-30 (Reprinted in : Feigenbaum, E.A., and J.Feldman, eds, Computers and Thought, pp. 406-450, New York ; McGraw-Hill.)

50. Minsky, M.L., 1967, *Computation ; Finite and Infinite Machines*, Englewood Cliffs, NJ ; Prentice-Hall.

51. Gabor, D., 1954, ''*Communication theory and cybernetics*'', IRE Transactions on Circuit Theory, vol. CT-1, pp. 19-31.

52. Taylor, W.K., 1956, ‘’Electrical simulation of some nervous system functional activities’’, Information Theory, vol. 3, E.C. Cherry, ed., pp. 314-328, London ; Butterworths.

53. Steinbuch, K., 1961, ‘’Die Lernmatrix ‘’; Kybernetik, vol., PP. 36-45.

54. Anderson, J.A., 1972 ; ‘’A simple neural network generating an interactive memory’’, Mathematical Biosciences, vol. 14, pp. 197-220.

55. Kohonen, T., 1972, ‘’Correlation matrix memories’’, IEEE Transactions on Computers, vol. C-21, pp. 353-359.

56. Nakato, K., 1972, ‘’Association – a model of associative memory’’, IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-2, pp. 380-388.

57. von Neumann, J., 1956, ‘’Probabilistic logics and the synthesis of reliable organisms from unreliable components’’ in Auromata Studies, C.E. Shannon and J. MacCarthy, eds., pp. 43-98, Princeton, NJ ; Princeton University Press.

58. Winograd, S., and J.D. Cowan ; 1963 ; Reliable Computation in the Presence of Noise, Cambridge, MA ; MIT Press.

59. Rosenblatt, F., 1958, ‘’The Perceptron : A probabilistic model for information storage and organization in the brain’’, Psychological Review, vol. 65. Pp. 386- 408.

60. Widrow, B., 1962, ‘’Generalization and information storage in networks of adeline ‘neurons’,’’ in M.C. Yovitz, G.T. Jacobi, and G.D., Goldstein, eds., Self-Organizing Systems, pp. 435-461, Washington, DC :Spartan Books.

61. Amari, S., 1967, ‘’A theory of adaptive patter, classifiers’’, IEEE Trans. Electronic Computers, vol. EC-16, pp.299-307.

62. Minsky, M.L., and S.A. Papert, 1969 , Perceptrons, Cambridge, MA ; MIT Press.

63. von der Malsburg, C., 1973, ‘’Self-organization of orientation sensitive cells in the striate cortex’’, Kybernetik, vol. 14 , pp. 85-100.

64. Willshaw, D.J., and C. von der Malsburg, 1976, ‘’How patterned neural connections can be set up by self-organization’’, Proceedins of the Ryal Society of London Series B, vol. 194, pp. 431-445.

65. Grossberg, S., 1980, ‘’How does a brain build a cognitive code ?’’, Psychological Review, vol.87, pp. 1-51.

66. Grossberg, S., 1972, ‘’Neural expectation :Cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes’’, Kybernetik, vol. 10, pp. 49-57.

67. Grossberg, S., 1976a, ‘’Adaptive pattern classification and universal recoding : I. Parallel development and coding of neural detectors’’, Biological Cybernetics, vol. 23, pp. 121-134.

68. Grossberg, S., 1976b, ‘’Adaptive pattern classification and universal recoding : II. Feedback, expectation, olfaction, illusions.’’, Biological Cybernetics, vol. 23, pp. 187-202.

69. Hopfield, J.J., 1982, ‘’Neural networks and physical systems with emergent collective computational abilities’’, Proceedings of the National Academy of Sciences, USA, vol. 79, pp. 2554-2558.

70. Cohen, M.A., and S., Grossberg, 1983, ''Absolute stability of global pattern formation and parallel memory storage by competitive neural networks, '' IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-13, pp. 815-826.

71. Kohonen, T., 1982, ''Self-organized formation of topologically correct feature maps'', Biological Cybernetics, vol.43, pp. 59-69.

72. Ackley, D.H., G.E. Hinton, and T.J. Sejnowski, 1985, '' A Learning Algorithm for Boltzmann Machines'', Cognitive Science, vol. 9, pp. 147- 169.

73. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, 1986a, ''Learning representations of back-propagation errors'', Nature (London), vol. 323, pp. 533-536.

74. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, 1986b, ''Learning internal representations by error propagation'', in D.E ; Rumelhart and J.L. McCleland, eds., vol 1, Chapter 8, Cambridge, MA ; MIT Press.

75. Rumelhart, D.E. and J.L. McCleland, eds., 1986, Parallel Distributed Processing ; Explorations in the Microstructure of Cognition, vol. 1, Cambridge, MA ; MIT Press.

76. Linsker, R., 1988b, ''Towards an organizing principle for a layered perceptual network'', in Neural Information Processing Systems, D.Z. Anderson, ed., pp.485-494, New York ; American Institutz of Physics.

77. Broomhead, D.S., and D. Lowe, 1988, ''Multivariate functional interpolation and adaptive networks'', Complex Systems, vol.2, pp.321-355.

78. Poggio, T., and F. Girosi, 1990b, ''Regularization algorithms for learning that are equivalent to multilayered networks'', Science, vol. 247, pp. 978-982.

79. Vapnik, V.N., 1992, ''Principles of risk minimization for learning theory'', Advances in Neural Information Processing Systems, vol. 4, pp. 831-838, San Mateo, CA ;Morgan Kaufmann.

80. Vapnik, V.N., 1995, ''The nature of Statistical Learning Theory'', New York ;Springer-Verlag.

81. Vapnik, V.N., 1998, 'Statistical Learning Theory'', New York ; Wiley.

82. C. Henssge, B. Knight, T. Krompecher, B. Madea, L. Nokes , (2002), The estimation of the time since death in the early postmortem period, Arnold, London, 2nd edn.

83. MUNOZ-BARUS José I. ; RODRIGUEZ-CALVO Maria Sol ; SUAREZ-PENARANDA José M. VIEIRA Duarte N. ; CADARSO-SUAREZ Carmen ; FEBRERO-BANDE Manuel, (2010) ; 'PMICALC: An R code-based software for estimating post-mortem interval (PMI) compatible with Windows, Mac and Linux operating systems' ; Forensic Science International, vol. 194, no1-3, pp. 49-52 ;

84. Terrence L. Fine, (1999), Feedforward Neural Network Methodology ; Springer – Verlag New York.

85. Simon Haykin, (1999), Neural Networks, a comprehensive foundation 2e edition, Prnetice-Hall, Inc .

86. Christopher M. Bishop ; (1995), Neural Networks for Pattern Recognition ; Clarendon Press, Oxford.

87. Robert A. Dunne ; (2007), A Statistical Approach to Neural Networks for Pattern Recognition ;; Wiley-Interscience.

88. Davia Lucy ; (2005), Introduction to Statistics for Forensic Scientists ; John Wiley &Sons Ltd .

89. Don Hong, Yu Shyr ; (2007), Quantitative medical data analysis using mathematical tools ans statistical techniques ; World Scientific Publishing .

90. Hagan, M.T., and M. Menhaj, (1994), "Training feed-forward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, Vol. 5, No. 6, 1999, pp. 989-993.

91. Moller, M.F., (1993), "A scaled conjugate gradient algorithm for fast supervised learning," Neural Networks, Vol. 6, pp. 525-533.

92. Hornik K. Stinchcombe M., White H. ; 1989, Multilayer feeforward networks are universal approximators, Neural Networks, 2, pp 359-366

93. Hornik K. Stinchcombe M., White H. ; 1990, Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, Neural Networks, 3 pp551-560.

94. Hornik K.,1991, Approximations capabilities of multilayer feedforward networks, Neural Networks, 4, pp 251-257.

95. Levenberg K., 1944, A method for the solution of certain nonlinear problems in least squares, Quaterly Journal of Applied Mathematics, 2, pp 164-168.

96. Marquardt D.W., 1963, An algorithm for least-squares estimation of non linear parameters, journal of the Society of Industrial and Applied Mathematics, 11, pp 431-441.

97. Sandhya Samarasinghe, 2007, ‘’Neural Networks for Applied Sciences and Engineering ; From Fundamentals to Complex Pattern Recognition’’, Auerbach Publications, eds. , New York

98. G.Dreyfus, J.-M. Martinez, M.Samuelides, M.B. Gordon, F. Badran, S.Thiria, L. Hérault ; 2002, ‘’Réseaux de neurones ; Méthodologie et applications’’, Editions Eyrolles, Paris.

99. www.pathguy.com/TimeDead.htm , Ed Friedlander MD.

100. www.swisswuff.ch/calculators/todeszeit.php , Wolf Schweitzer, MD, Institute of Legal Medicine, University of Zuerich, Switzerland.

101. Ingo Steinwart, Andreas Christmann, 2008, ‘’Support Vector Machines’’, John Wiley &Sons, Ltd ; New York.

TABLE DE MATIERE