

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ KASDI MERBAH, OUARGLA
Faculté des Mathématiques et des Sciences de la Matière
DÉPARTEMENT DE MATHÉMATIQUES



Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : **Probabilités et statistique**

Par

souhir labeled

Titre :

**La régression sous le modèle neuronal et
classique**

Membres du Comité d'Examen :

Kouadri ghada **M.C.B** Université Ouargla **Président**
Agoune Rachid **M.A.A** Université Ouargla **Encadreur**
Abassi hocine **M.A.A** Université Ouargla **Examineur**

Juin 2023

Dédicace

A l'occasion de mon diplôme, de ma réussite, et de tout ce qui est beau dans ma
vie, à ma famille, car ils sont la source de ma force et de ma réussite
Maintenant, je me sens si reconnaissant et fier que vous soyez mes parents, merci
moi-même

A mes chères sœurs : Ikram, ouissal, bassair, hadil

A mon frère :Ishak

Je dédie ce travail à mon modèle :Prof REZIG ABD ELDJEBBAR

Aux compagnons de ma vie : Asma, Ahlam

A mes amis : imane

A toute ma famille labed, source d'espoir et de motivation

Je dédie ce travail à ...

SOUHIR LABED © 2023

REMERCIEMENTS

Je remercie d'abord et avant tout le bon Dieu qui m'a donné la foi, le courage, la
volonté et la patience pour réaliser ce travail.

Je tiens à remercier mon encadreur M. Agoune Rachid pour son aide précieuse
ainsi

que les ses précieux conseils et directives qui m'ont permis d'accomplir ce travail de
recherche.

Mes vifs remerciements sont également adressés aux membres du jury Dr. Kouadri
ghada

et Dr. Abassi hocine qui ont accepté d'évaluer ce travail.

Ainsi je remercie tous les enseignants de la faculté des mathématiques et des sciences de la matière du département de mathématique. Enfin un merci particulier à tous mes camarades sans exception.

SOUHIR LABED © 2023

Table des matières

Remerciements	1
Table des matières	3
Table des figures	6
1 Généralités	9
1.1 L'espérance mathématique	9
1.2 la variance	9
1.3 La régression	10
1.4 La méthode des moindres carrés ordinaire (MCO)	11
1.5 Equation d'analyse de la variance	11
1.5.1 SCT	11
1.5.2 SCE	11
1.5.3 SCR	11
1.6 Fonction d'activation	11
1.6.1 Fonction de Heaviside (seuil)	11
1.6.2 Fonction de signe	12
1.6.3 Fonction linéaire	12

1.6.4	Fonction linéaire à seuil ou multi-seuil	12
1.6.5	Fonction de densité de la loi normal	12
1.6.6	La fonction tangente hyperbolique	12
1.6.7	La fonction exponentielle	13
2	Le modèle de régression linéaire multiple	14
2.1	Représentation du modèle	14
2.2	Estimateur des moindres carrés ordinaires	16
2.2.1	Estimateur sans biais	17
2.2.2	Estimateur de la matrice de variance covariance des	18
2.3	Equation d'analyse de la variance et qualité d'un ajustement	19
2.4	Estimation par intervalles de confiance et tests d'hypothèses asymp-	
	totiques	20
2.4.1	Théorèmes limites	20
2.5	Les tests statistiques	21
2.5.1	Le test de student	21
2.5.2	Test de Fisher (test de signification globale du modèle de re-	
	gression)	22
3	Les réseaux de neurones	23
3.1	Neurone formel	23
3.2	Fonction d'activation	25
3.2.1	Fonction gaussienne	25
3.2.2	La fonction sigmoïde	26
3.3	Propriétés importantes de la fonction d'activation	27
3.4	Les réseaux de neurones formels	28

3.4.1	Le réseau monocouche	29
3.5	Les réseaux de neurones bouclés et non bouclés	29
3.5.1	Les réseaux non bouclés	29
3.5.2	Réseaux à couches	30
3.5.3	Forme générale	30
3.6	Réseaux à une couche cachée de sigmoïdes et un neurone de sortie	
	linéaire	31
3.7	L'estimateur de régression multiple par réseau de neurones	32
3.8	la descente de gradient	33
	Annexe B : Abréviations et Notations	42

Table des figures

3.1 Un neurone réalise une fonction non linéaire bornée	24
3.2 Neurone formel	25
3.3 la fonction sigmoïde	26
3.4 Le réseau mon couche	28
3.5 réseaux multicouches	29
3.6 Un éseau de neurones à n entrées, une couche de N_c neurones cachés,et N_c neurones de sortie	31

Introduction

L'une des méthodes statistiques les plus utilisées dans divers domaines de la science est la méthode d'analyse de régression, elle explique la relation entre les variables sous la forme d'une relation exprimée par une équation basée sur l'estimation de ses coefficients, et que cette question est d'une grande importance dans la planification et la prise des décisions. De nombreuses méthodes d'estimation des coefficients de ce modèle ont été développées dont les plus importantes est la méthode moindres carrés ordinaires et la méthode basé sur les réseaux de neurones par la technique de la descente de gradient.

Des neurones artificiels auxquels les chercheurs sont intéressés ces dernières années. Les réseaux de neurones artificiels sont l'un des domaines les plus importants de l'intelligence artificielle, qui sert à représenter l'esprit humain à travers un réseau de données numériques (qui représente le neurone) connecté à des liens pondérés (qui représentent les connexions entre les cellules) et qui est traité par un modèle mathématique spécifique. Les données seules n'ont aucun effet, mais si ces données sont unies aux poids, elles conduisent à déterminer la tâche requise du réseau de neurones. Cette méthode a été active dans de nombreux domaines statistiques, y compris l'estimation des coefficients des modèles de régression. Dans ce mémoire, l'objectif de l'étude était de comparer les deux méthodes d'estimation de régression multiple linéaire les moindres carrées ordinaires et les réseaux de neurones artificiels en choisissant le meilleur dans l'estimation des modèles en utilisant le critère d'erreur

quadratique moyenne.

Chapitre 1

Généralités

1.1 l'espérance mathématique

Définition 1.1.1 *l'espérance mathématique d'une variable aléatoire réelle est, intuitivement, la valeur que l'on s'attend à trouver, en moyenne, si l'on répète un grand nombre de fois la même expérience aléatoire. Elle se note $\mathbb{E}[X]$ et se lit « espérance de X ».*[\[1\]](#)

1.2 la variance

Définition 1.2.1 *est une mesure de la dispersion d'un échantillon d'une variable aléatoire. Elle exprime la moyenne des carrés des écarts à la moyenne, aussi égale à la différence entre la moyenne des carrés des valeurs de la variable et le carré de la moyenne, selon le théorème de König-Huygens. Ainsi, plus l'écart à la moyenne est grand plus il est prépondérant dans le calcul total (voir la fonction carré) de la variance qui donnerait donc une bonne idée sur la dispersion des valeurs. La variance est toujours positive, et ne s'annule que s'il n'y a essentiellement qu'une seule valeur.*

La racine carrée définit l'écart type σ , d'où la notation [2]

$$\sigma^2 = v[X] = \text{var}[X] \quad (1.1)$$

1.3 La régression

Définition 1.3.1 *La régression est une méthode statistique couramment utilisée pour explorer la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Elle permet de déterminer si une relation existe entre ces variables, de la quantifier et de formuler des prévisions ou des estimations sur la base de cette relation.*

La régression peut être utilisée pour diverses applications telles que :

- *Prédire une variable dépendante à partir d'une ou plusieurs variables indépendantes, par exemple, prédire le prix des maisons en fonction de leur taille, de leur emplacement, etc [7].*
- *Etudier l'impact d'une variable indépendante sur une variable dépendante, par exemple, évaluer le niveau d'influence du niveau d'éducation sur le salaire.*
- *Comprendre et modéliser la relation entre plusieurs variables indépendantes et une variable dépendante.*

Il existe différents types de modèles de régression, tels que la régression linéaire, la régression logistique, la régression polynomiale. Le choix du modèle dépendra du type de données et de la relation entre les variables étudiées. La régression est largement utilisée dans divers domaines tels que l'économie, les sciences sociales, les sciences naturelles, la finance, la recherche marketing, etc [7]

la régression linéaire multiple : est une généralisation de la régression linéaire simple, dans le sens où cette approche permet d'évaluer les relations linéaires entre une variable réponse et plusieurs variables explicatives (de type numérique ou catégoriel).

1.4 La méthode des moindres carrés ordinaire (MCO)

est le nom technique de la régression mathématique en statistiques, et plus particulièrement de la régression linéaire. Il s'agit d'un modèle couramment utilisé en économétrie.

1.5 Equation d'analyse de la variance

1.5.1 SCT

est la somme des carrés des écarts totale, c'est la variabilité totale de Y

1.5.2 SCE

est la somme des carrés des écarts expliquée par l'ajustement affine, c'est la variabilité de Y expliquée par l'ajustement affine.

1.5.3 SCR

est la somme des carrés des écarts résiduelle, c'est la variabilité résiduelle (inexpliquée) de Y .

1.6 Fonction d'activation

1.6.1 Fonction de Heaviside (seuil)

$$S = \begin{cases} 1 & \text{si } \alpha \geq 0 \\ 0 & \text{si } \alpha < 0 \end{cases} \quad (1.2)$$

1.6.2 Fonction de signe

$$S = \begin{cases} 1 & \text{si } \alpha \geq 0 \\ -1 & \text{si } \alpha < 0 \end{cases} \quad (1.3)$$

1.6.3 Fonction linéaire

$$S = \lambda\alpha$$

1.6.4 Fonction linéaire à seuil ou multi-seuil

$$F(x) = \begin{cases} x & x \in [u, v] \\ v & \text{si } x \geq v \\ u & \text{si } x \leq u \end{cases} \quad (1.4)$$

1.6.5 Fonction de densité de la loi normal

$$s = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \quad (1.5)$$

1.6.6 La fonction tangente hyperbolique

La fonction tangente hyperbolique, définie par :

$$\frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (1.6)$$

Est aussi très utilisée en pratique, car elle partage avec la fonction sigmoïde certaines caractéristiques pratiques :

- non polynômiale
- indéfiniment continûment dérivable
- calcul rapide de la dérivée par la formule

$$\frac{d}{dx}th(x) = 1 - (th(x))^2 \quad (1.7)$$

1.6.7 La fonction exponentielle

Elle est souvent utilisée au niveau de la couche de sortie afin d'assurer des valeurs de sorties positives et non bornées. L'utilisation des fonctions de transfert non-linéaires permet l'obtention de modèles statistiques non-linéaires.

Chapitre 2

Le modèle de régression linéaire multiple

La régression linéaire multiple est une généralisation de la régression linéaire simple [5]. Son but est d'étudier et modéliser la relation entre une variable expliquée Y et plusieurs variables explicatives X_1, X_2, \dots, X_p

2.1 Représentation du modèle

Le modèle de régression linéaire multiple s'écrit sous la forme

$$Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \text{ pour } i = 1 \dots n \quad (2.1)$$

où,

$$p \leq n$$

y_i est la $i^{\text{ème}}$ observation de Y

x_i la $i^{\text{ème}}$ observation de $(x_1, x_2, \dots, x_{p-1})$

$(\beta_0, \beta_1, \dots, \beta_{p-1})$ les paramètres à estimer (coefficients de régression)

les ε_i sont des variables aléatoires, non observées, appelées erreurs ou bruits, auxquelles on impose certaines conditions complémentaires.

les hypothèses relatives à ce modèle sont :

- ▶ $\forall i \neq j, cov(\varepsilon_i, \varepsilon_j) = 0.$
- ▶ $\forall i, E[\varepsilon_i] = 0.$
- ▶ $var[\varepsilon_i] = \sigma^2.$
- ▶ $cov(\varepsilon_i, x_i) = 0.$

Ce modèle est appelé modèle de régression linéaire multiple. Il s'exprime de façon vectorielle :

$$Y = X\beta + \varepsilon \quad (2.2)$$

avec

$$Y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{1,0} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,0} & \dots & x_{n,p} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \cdot \\ \cdot \\ \cdot \\ \beta_P \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix} \quad (2.3)$$

Sous les conditions, on a alors :

- $E[\varepsilon] = 0$
- $E[Y] = X\beta$
- $var[\varepsilon] = var[Y] = \sigma^2 I_n$
- en absence de colinéarité entre les variables explicatives, i.e. $X'X$ est régulière,

$$\det(X'X) \neq 0$$

et $(X'X)^{-1}$ existe

2.2 Estimateur des moindres carrés ordinaires

Afin d'estimer le vecteur β composé des coefficients $(\beta_0, \beta_1, \dots, \beta_p)$, nous appliquons la méthode des moindres carrés ordinaire (MCO) qui consiste à minimiser la somme des carrés des erreurs [8], soit :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^n \left(Y_i - \sum_{j=0}^{p-1} \beta_j x_{i,j} \right)^2 = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \| Y - X\beta \|^2, \quad (2.4)$$

$\| \cdot \|$ est la norme euclidienne de \mathbb{R}^n ,

on montre que

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (2.5)$$

soit

$$\begin{aligned} L : \beta &\mapsto (Y - X\beta)'(Y - X\beta) = Y'Y - Y'X\beta - X'\beta'Y + X'\beta'X\beta \\ &= Y'Y - 2X'\beta'Y + X'\beta'X\beta \end{aligned} \quad (2.6)$$

Le vecteur

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (2.7)$$

est bien un point critique de L puisque

$$\nabla L(\hat{\beta}) = 2X'X\hat{\beta} - 2X'Y \quad (2.8)$$

Ce point critique correspond à un minimum. En effet, la matrice hessienne de L en $\hat{\beta}$ vaut $2X'X$ qui est définie positive. On introduit maintenant, comme pour la régression linéaire simple, le sous-espace vectoriel $\varepsilon(X)$ de \mathbb{R}^n engendré par les vecteurs colonnes de X . Par définition $X\hat{\beta}$ est un vecteur de $\varepsilon(X)$ dont la distance euclidienne avec Y est la distance minimum entre Y et tout vecteur de $\varepsilon(X)$. Par

conséquent, si l'on note Π_X la matrice de projection orthogonale sur $\varepsilon(X)$ alors

$$X\hat{\beta} = \Pi_X Y$$

Là encore, on peut montrer que la matrice Π_X s'écrit aussi

$$X(X'X)^{-1}X'$$

d'où

$$X\hat{\beta} = X(X'X)^{-1}X'Y$$

puis

$$\hat{\beta} = (X'X)^{-1}X'Y$$

2.2.1 Estimateur sans biais

Soit le modèle [\[3\]](#) :

$$Y = X\beta + \varepsilon$$

On peut s'écrire :

$$\begin{cases} Y = X\beta + \varepsilon \\ \hat{Y} = X\hat{\beta} \end{cases} \Rightarrow \varepsilon = Y - \hat{Y} \quad (2.9)$$

Nous obtenons :

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y & (2.10) \\ &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= (X'X)^{-1}X'(X\beta) + (X'X)^{-1}X'\varepsilon \\ &= \beta + (X'X)^{-1}X'\varepsilon \end{aligned}$$

$$\hat{\beta} - \beta = (X'X)^{-1}X'\varepsilon \quad (2.11)$$

D'où

$$E[\hat{\beta}] = \beta + (X'X)^{-1}X'E[\varepsilon] \quad (2.12)$$

avec :

$$E[\varepsilon] = 0$$

Finalement

$$E[\hat{\beta}] = \beta$$

l'estimateur est sans biais.

2.2.2 Estimateur de la matrice de variance covariance des

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1} \quad (2.13)$$

Preuve. en utilisant l'estimateur sans biais :

$$\begin{aligned} Var((X'X)^{-1}X'Y) &= (X'X)^{-1}X'Var(Y)X(X'X)^{-1} \\ &= (X'X)^{-1}X'(\sigma^2 I_s)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \quad (2.14)$$

Avec σ^2 matrice diagonale :

$$\begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2 \end{pmatrix}$$

Et puisque σ^2 est inconnu donc on l'estime. ■

2.3 Equation d'analyse de la variance et qualité d'un ajustement

$$SCR = \|\hat{\varepsilon}\|^2$$

$$SCE = \|\hat{Y} - \bar{Y}\|^2$$

$$SCT = \|Y - \bar{Y}\|^2$$

par le théorème de Pythagore :

$$SCT = SCE + SCR \tag{2.15}$$

(équation d'analyse de la variance)

$$\|Y - \bar{Y}\|^2 = \|\hat{Y} - \bar{Y}\|^2 + \|\hat{\varepsilon}\|^2 \tag{2.16}$$

$$SCT = SCE + SCR$$

avec :

SCT : variabilité totale

SCE : variabilité expliquée

SCR : variabilité résiduelle.

Cette équation permet de juger la qualité d'ajustement d'un modèle, en effet plus SCE est proche du SCT meilleur est l'ajustement globale du modèle. Cependant ces valeurs dépendent des unités de mesure, c'est pourquoi on préfère utiliser le nombre sans dimensions :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \tag{2.17}$$

R^2 = Coefficient de détermination : mesure la proportion de la variance de Y expli-

quée par la régression de Y sur X .

◆ Si $n < k$ alors on calcule le coefficient de détermination corrigé

$$\bar{R}^2 = 1 - \left[\frac{n-1}{n-k-1} (1 - R^2) \right] \quad (2.18)$$

on a $\bar{R}^2 \leq R^2$ et si n est grand $\bar{R}^2 \simeq R^2$

2.4 Estimation par intervalles de confiance et tests d'hypothèses asymptotiques

Pour construire des intervalles de confiance ou des tests d'hypothèses sur β , on a besoin de connaître la loi de $(\hat{\beta} - \beta)$. Dans le cas général, on ne fait aucune hypothèse sur la loi de ε , donc on peut éventuellement chercher à en avoir une connaissance approximative lorsque n est très grand. [\[4\]](#)

2.4.1 Théorèmes limites

Pour chaque taille d'échantillon n , on précisera la dépendance en n à l'aide de (n) : Y sera ainsi noté $Y^{(n)}$, X , $\hat{\beta}$, $\hat{\varepsilon}$ et $\hat{\sigma}^2$ deviennent respectivement $X^{(n)}$, $\hat{\beta}^{(n)}$, $\hat{\varepsilon}^{(n)}$ et $\hat{\sigma}^{2(n)}$ si les ε_i sont maintenant supposées *i.i.d.* et si

$$\frac{X^{(n)'} X^{(n)}}{n} \rightarrow A \quad (2.19)$$

définie positive lorsque $n \rightarrow +\infty$ alors

$$\sqrt{n}(\hat{\beta}^{(n)} - \hat{\beta}) \xrightarrow{(\mathcal{L})} N(0; \sigma^2 A^{-1}) \quad (2.20)$$

Remarque 2.4.1 σ^2 étant inconnu, on l'estime par $\hat{\sigma}^{2(n)}$.

2.5 Les tests statistiques

2.5.1 Le test de student

Tester l'influence directe de la variable explicative sur la variable endogène, revient à tester son coefficient de régression s'il est égale ou différent de 0, pour un seuil choisi, en général $\alpha = 0.05$, Le test d'hypothèse est le suivant : est appelé (test bilatéral)

$$\begin{cases} H_0 : \beta_i = 0 & , i = 1, \dots, p \\ H_1 : \beta_i \neq 0 \end{cases}$$

La statistique de student est la suivante :

$$T_c^\alpha = \left| \frac{\hat{\beta}_i - \beta_i}{\sigma(\hat{\beta})} \right| \rightarrow T_{(n-k-1, \frac{\alpha}{2})} \quad (2.21)$$

Règle de décision

Si

$$|T_c| \leq T_t^{\alpha=0,05} \quad (2.22)$$

on accepte l'hypothèse H_0 : la variable x_i n'est pas contributive à l'explication de Y .

Test unilatéral : ce test est utilisé lorsque $H_1 : \beta_i > 0$, ou $\beta_i < 0$

$$T_c^\alpha = \left| \frac{\hat{\beta}_i - \beta_i}{\sigma(\hat{\beta})} \right| \rightarrow T_{(n-k-1, \alpha)} \quad (2.23)$$

2.5.2 Test de Fisher (test de signification globale du modèle de regression)

Pour tester si l'ensemble des variables explicatives ont une influence sur la variable à expliquée, on fait le test d'hypothèse suivant :

$$\begin{cases} H_0 : \beta_0, \beta_1, \dots, \beta_k = 0 \\ H_1 : \text{il existe au moins, } \beta_i \neq 0 \text{ pour } i = 1, \dots, p \end{cases} \quad (2.24)$$

A partir de l'équation de l'analyse de la variance on a

$$\begin{aligned} \sum_{i=1}^n (Y_i - \hat{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \varepsilon_i^2 \\ &= E \left(\sum_{i=1}^n \varepsilon_i^2 \right) \\ &= (n - k - 1) \sigma^2 \end{aligned} \quad (2.25)$$

Chapitre 3

Les réseaux de neurones

Les réseaux de neurones formels sont à l'origine d'une tentative de modélisation mathématique du cerveau humain. Les premiers travaux datent de 1943 et sont l'oeuvre de Mac Culloch et Pitts. Ils présentent un modèle assez simple pour les neurones et explorent les possibilités de ce modèle. La modélisation consiste à mettre en oeuvre un système de réseau neuronal sous un aspect non pas biologique mais formel, cela suppose que d'après le principe biologique on aura une correspondance pour chaque élément composant le neurone biologique, donc une modélisation pour chacun d'entre eux [6]

3.1 Neurone formel

Le modèle de neurone formel dû à Mac Culloch et Pitts la figure (3.1) est un processeur élémentaire. Il reçoit un nombre variable d'entrées en provenance de neurones amont.

A chacune de ces entrées est associée un poids W , représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie ensuite pour alimenter, un nombre variable de neurones. A chaque connexion est

associé un poids. Un neurone est une fonction non linéaire, paramétrée, à valeurs bornées. Les variables sur lesquelles opère le neurone sont habituellement désignées sous le terme d'entrées du neurone, et la valeur de la fonction sous celui de sortie. Il est commode de représenter graphiquement un neurone comme indiqué sur la figure (3.1). Le neurone est défini par trois caractéristiques : son état, ses connexions avec d'autres neurones et sa fonction de transfert. Nous utiliserons les notations suivantes : O : l'ensemble des états possibles des neurones.

o_i : l'état du neurone i , $o_i \in O$.

f_i : la fonction de transfert (d'activation) associée au neurone i .

s_i : l'entrée du neurone i , $s_i \in \mathbb{R}$.

w_{ij} : le poids de la connexion du neurone j vers le neurone i , $w_{ij} \in W$ l'ensemble des poids du *PMC*

$$y = f(x_1, x_2, \dots, x_n; w_1, w_2, \dots, w_p) \tag{3.1}$$

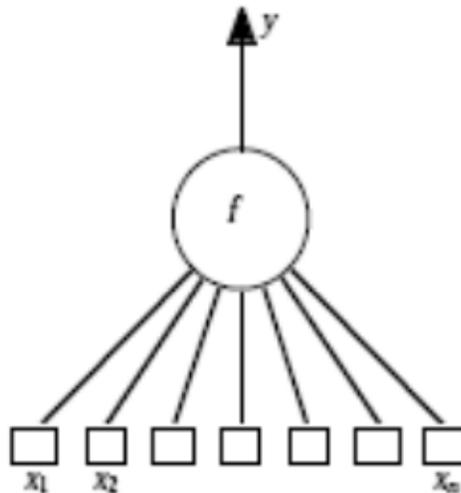


FIG. 3.1 – Un neurone réalise une fonction non linéaire bornée

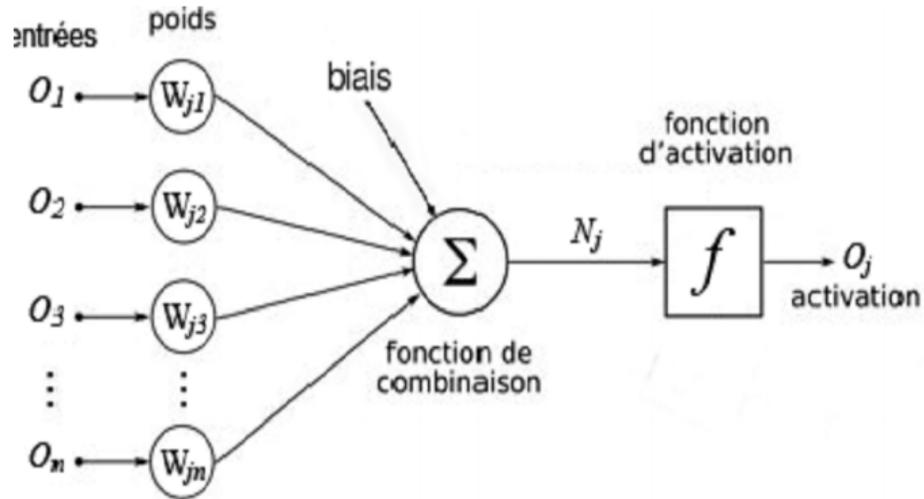


FIG. 3.2 – Neurone formel

3.2 Fonction d'activation

La fonction d'activation (la fonction de transfert) joue un role très important dans le comportement du neurone. Elle retourne une valeur représentative de l'activation du neurone, cette fonction, a comme paramètre la somme pondérée des entrées ainsi que le seuil d'activation. La nature de cette fonction diffère selon le réseau. On en compte divers types, parmi elles :

3.2.1 Fonction gaussienne

$$s = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \quad (3.2)$$

3.2.2 La fonction sigmoïde

est la plus utilisée car elle introduit la non-linéarité, mais c'est aussi une fonction continue, différentiable. Une fonction sigmoïde est définie par :

$$f_{sig}(x) = \frac{1}{1 + e^{-x}} \quad (3.3)$$

Possède les propriétés importantes évoquées précédemment (elle n'est pas polynomiale et est indéfiniment continûment dérivable). En outre, une propriété simple permet d'accélérer le calcul de sa dérivée, ce qui réduit le temps de calcul nécessaire à l'apprentissage d'un réseau de neurones. On a en effet

$$\frac{d}{dx} f_{sig}(x) = f_{sig}(x) (1 - f_{sig}(x)) \quad (3.4)$$

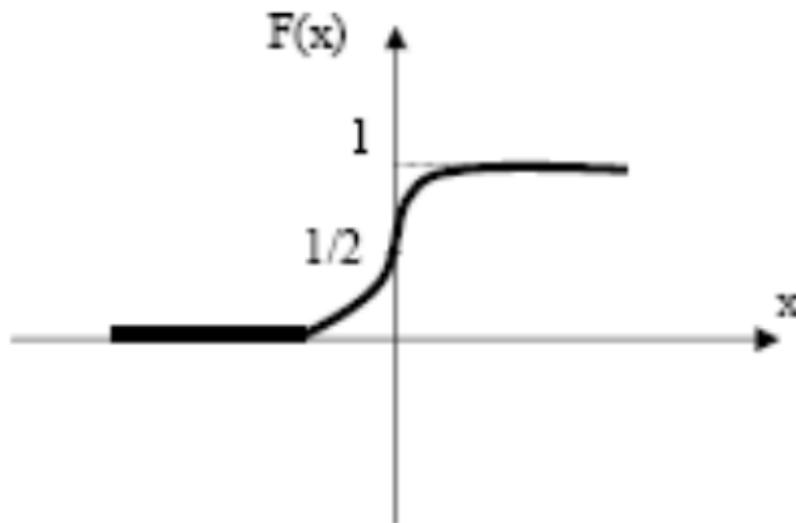


FIG. 3.3 – la fonction sigmoïde

On peut donc calculer la dérivée de cette fonction en un point de façon très efficace

à partir de sa valeur en ce point.

De plus, la fonction sigmoïde est à valeurs dans l'intervalle $[0; 1]$, ce qui permet d'interpréter la sortie du neurone comme une probabilité.

3.3 Propriétés importantes de la fonction d'activation

Les propriétés de la fonction d'activation influent en effet sur celle du neurone formel et il est donc important de bien choisir celle-ci pour obtenir un modèle utile en pratique.

Quand les neurones sont combinés en un réseau de neurones formels, il est important par exemple que la fonction d'activation de certains d'entre eux ne soit pas un polynôme sous réserve de limiter la puissance de calcul du réseau obtenu. Un cas caricatural de puissance limitée correspond à l'utilisation d'une fonction d'activation linéaire, comme la fonction identité : dans une telle situation le calcul global réalisé par le réseau est lui-aussi linéaire et il est donc parfaitement inutile d'utiliser plusieurs neurones, un seul donnant des résultats strictement équivalents.

Cependant, les fonctions de type sigmoïde sont généralement bornées. Dans certaines applications, il est important que les sorties du réseau de neurones ne soient pas limitées a priori : certains neurones du réseau doivent alors utiliser une fonction d'activation non bornée. On choisit généralement la fonction identité. Il est aussi utile en pratique que la fonction d'activation présente une certaine forme de régularité. Pour calculer le gradient de l'erreur commise par un réseau de neurones, lors de son apprentissage, il faut que la fonction d'activation soit dérivable. Pour calculer l'erreur " la matrice hessienne de l'erreur ", ce qui est utile pour certaines analyses d'erreur, il faut que la fonction d'activation soit dérivable deux fois. Comme elles

comportent généralement des points singuliers, les fonctions linéaires par morceaux sont relativement peu utilisées en pratique. [6]

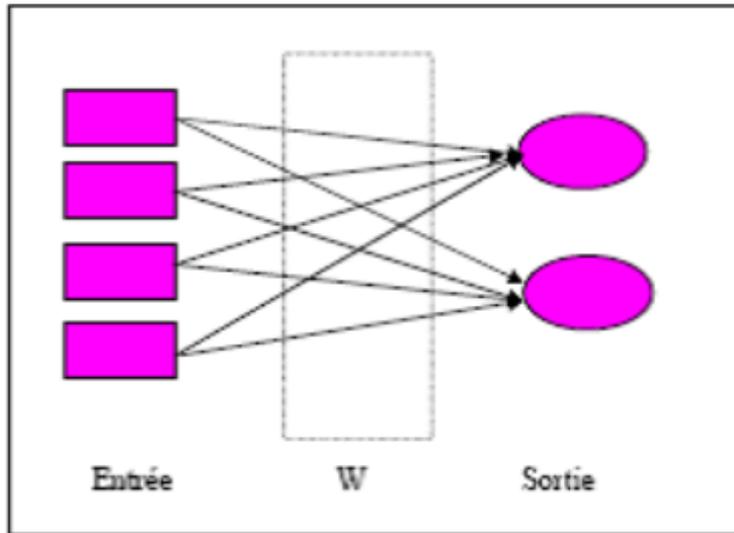


FIG. 3.4 – Le réseau mon couche

3.4 Les réseaux de neurones formels

Un réseau de neurones peut être défini comme un graphe orienté et pondéré. Les neurones étant les noeuds, qui seront connectés par des liens appelés liens synaptiques ou synapses.

Ces liens synaptiques sont pondérés par des poids judicieusement choisis. La propagation de l'activation à travers les liens synaptiques des neurones en aval va influencer les autres neurones en amont, cette activation sera pondérée par le lien qu'elle prendra ainsi on l'appellera poids synaptique.

3.4.1 Le réseau monocouche

La structure d'un réseau monocouche est telle que les neurones organisés en entrées soient entièrement connectés à d'autres neurones organisés en sortie par une couche modifiable de poids.

3.5 Les réseaux de neurones bouclés et non bouclés

3.5.1 Les réseaux non bouclés

Un réseau de neurone non bouclé est donc représenté graphiquement par un ensemble de neurones.

Un réseau de neurone non bouclé réalise une (ou plusieurs) fonctions de ses entrées, par composition des fonctions réalisées par chacun de neurones "Connectés" entre eux, l'information circulant des entrées vers les sorties sans "retour en arrière", si l'on représente le réseau comme

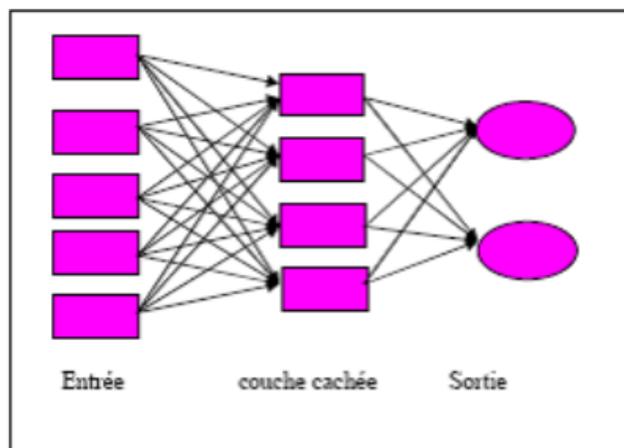


FIG. 3.5 – réseaux multicouches

un graphe dont les nœuds sont les neurones et les arêtes les "connexions" entre eux – ci,

le graphe d'un réseau non bouclé est non acyclique : si l'on se déplace dans le réseau, à partir d'un neurone quelconque, en suivant les connexions, on ne peut pas revenir au neurone de départ.

La représentation de la topologie d'un réseau par un graphe est très utile.

3.5.2 Réseaux à couches

La seule contrainte sur le graphe des connexions d'un réseau de neurones non bouclé est qu'il ne contienne pas de cycle. On peut donc imaginer une grande variété de topologies pour ces réseaux.

La très grande majorité des applications des réseaux de neurones mettent en jeu des " réseaux à couches ", dont un exemple est représenté sur la figure [\(3.5\)](#)

3.5.3 Forme générale

Ce réseau réalise N_o fonctions algébriques des variables d'entrées du réseau, chacune des sorties est une fonction, réalisée par le neurone de sortie correspondant, des fonctions non linéaires réalisées par les neurones cachés.

Un réseau de neurones non bouclé à n entrées, N_c neurones cachés et N_o neurones de sorties réalise N_o fonctions non linéaires de ses n variables d'entrées par composition des N_c fonctions réalisées par ses neurones cachés.

Remarque 3.5.1 *Remarque Le temps ne joue aucun rôle fonctionnel dans un réseau de neurones non bouclé : si les entrées sont constantes, les sorties le sont également. Le temps nécessaire pour le calcul de la fonction réalisée par chaque neurone est négligeable et, fonctionnellement, on peut considérer ce calcul comme instantané. Pour cette raison, les réseaux non bouclés sont souvent appelés " réseaux statiques ", par opposition aux réseaux bouclés ou " dynamiques " que nous introduirons plus bas.*

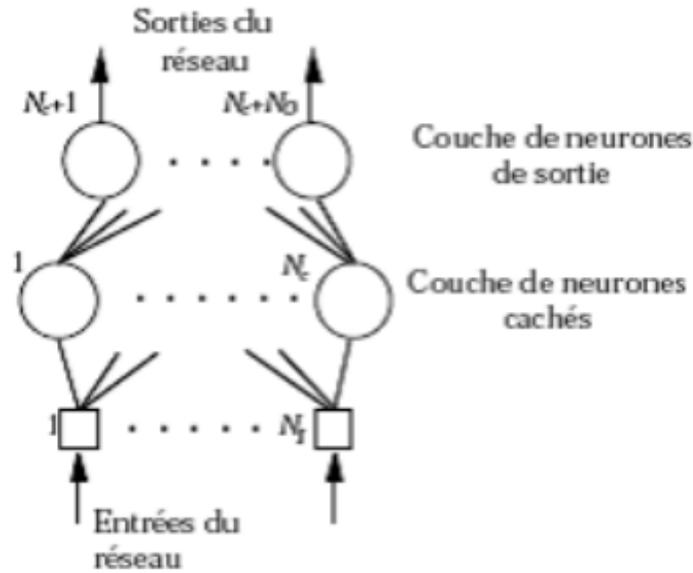


FIG. 3.6 – Un éseau de neurones à n entrées, une couche de N_c neurones cachés, et N_c neurones de sortie

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

Pour terminer cette présentation des réseaux de neurones non bouclés, considérons le cas, très important en pratique, des réseaux à une couche cachée à fonction d'activation sigmoïde et un neurone de sortie linéaire.

La sortie de ce réseau a pour expression :

$$g(x; w) = \sum_{i=1}^{N_c} \left[w_{N_c+1,i} \operatorname{th} \left(\sum_{j=1}^n w_{ij} x_j + w_{i0} \right) \right] + w_{N_c+1,0} = \sum_{i=1}^{N_c} \left[w_{N_c+1,i} \operatorname{th} \left(\sum_{j=0}^n w_{ij} x_j \right) \right] + w_{N_c+1,0} \quad (3.5)$$

Ou x est le vecteur des entrées (de dimension $n+1$), et w est le vecteur des paramètres, de dimension $(n+1)N_c + (N_c+1)$. 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 . La sortie du

réseau $g(x; w)$ est une fonction linéaire des poids de la dernière couche de connexions (qui relie le 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 relie les $n + 1$ entrées du réseau aux N_c neurones cachés). Cette propriété a des conséquences importantes que nous examinerons plus bas. La sortie d'un réseau de neurone non bouclé est une fonction non linéaire de ses entrées et de ses paramètres.

3.7 L'estimateur de régression multiple par réseau de neurones

L'estimateur de régression multiple par réseau de neurones est généralement basé sur une structure de réseau de neurones multicouche. La formule de l'estimateur de régression multiple par réseau de neurones peut varier en fonction de la structure et des paramètres du réseau de neurones utilisé. Cependant, en général, on peut définir que la formule de l'estimateur de régression multiple par réseau de neurones peut varier en fonction de la structure et des paramètres du réseau de neurones utilisé. Cependant, en général, on peut définir l'estimateur de régression multiple ainsi :

$$Y = f(X) = w_0 + w_1X_1 + w_2X_2 + \dots + w_mX_m \quad (3.6)$$

où Y est la variable à prédire, X est un vecteur de variables explicatives, w sont les poids du modèle, et f est la fonction d'activation utilisée par le réseau de neurones. La fonction d'activation permet de calculer une sortie en fonction de l'entrée du neurone. Il en existe plusieurs types, tels que la fonction sigmoïde, la fonction tangente hyperbolique, etc. La fonction d'activation utilisée dépendra principalement de la structure et des données du modèle.

La combinaison de différents types de couches (couches d'entrée, couches cachées et

couches de sortie) et de différentes fonctions d'activations permet au réseau de neurones de construire un modèle non linéaire qui peut capturer des relations complexes entre les variables explicatives et la variable à prédire.

Enfin, l'estimateur de régression multiple par réseau de neurones nécessite un processus d'apprentissage, également appelé entraînement, pour ajuster les poids du modèle afin de minimiser l'erreur de prédiction. Pour cela, on utilise généralement un algorithme d'optimisation tel que la descente de gradient.

3.8 la descente de gradient

- Version On-line : on considère à chaque fois seulement l'erreur d'une seule donnée en entrée

$$E = \frac{1}{2} (y(x^m, w) - y^m)^2 \quad (3.7)$$

- On suit la descente de gradient. Le gradient de l'erreur d'entraînement est calculé comme une fonction des poids. pour simplifier, on note $y(x_m, w)$ simplement par y .

$$\nabla_w E = (y(x^m, w) - y^m) \nabla_w y(x^m, w) \quad (3.8)$$

où

$$\nabla_w y(x^m, w) = \left[\frac{\partial y}{\partial w_1}, \frac{\partial y}{\partial w_2}, \dots, \frac{\partial y}{\partial w_n} \right]^t$$

On modifie w comme suit :

$$w \leftarrow w - \eta \nabla_w E \Leftrightarrow w \leftarrow w - \eta (y(x^m, w) - y^m) \left[\frac{\partial y}{\partial w_1}, \frac{\partial y}{\partial w_2}, \dots, \frac{\partial y}{\partial w_n} \right]^t$$

Gradient de la sortie

$$\begin{aligned} \nabla_w y(x^m, w) &= \left[\frac{\partial y}{\partial w_1}, \frac{\partial y}{\partial w_2}, \dots, \frac{\partial y}{\partial w_n} \right]^t \\ z &= \sum_{i=0}^n w_i x_i^m \quad y = s(z) = \frac{1}{1 + e^{-z}} \\ w &\leftarrow w - \eta (y(x^m, w) - y^m) \left[\frac{\partial y}{\partial w_1}, \frac{\partial y}{\partial w_2}, \dots, \frac{\partial y}{\partial w_n} \right]^t \end{aligned}$$

on a

$$\begin{aligned} \frac{\partial y}{\partial w_i} &= \frac{\partial y}{\partial z} \frac{\partial z}{\partial w_i} \\ &= \frac{\partial s(z)}{\partial z} \frac{\partial z}{\partial w_i} \\ &= \frac{\partial s(z)}{\partial z} x_i^m \end{aligned}$$

donc

$$w \leftarrow w - \eta (y - y^m) \frac{\partial s(z)}{\partial z} x_i^m$$

En notant :

$$\delta = \frac{\partial E}{\partial z} = (y - y^m) \frac{\partial s(z)}{\partial z} \tag{3.9}$$

on a : $\nabla w_i = -\eta \delta x_i^m$ Delta rule

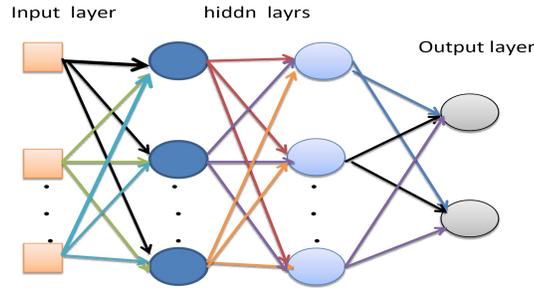
éivée de la sigmoïde :

$$\begin{aligned}
 s(z) &= \frac{1}{1 + e^{-z}} & (3.10) \\
 \frac{ds(z)}{dz} &= \frac{ds}{dz} [(1 + e^{-z})^{-1}] \\
 &= [-(1 + e^{-z})^{-2}] [-e^{-z}] \\
 &= \left[\frac{1}{1 + e^{-z}} \right] \left[\frac{e^{-z}}{1 + e^{-z}} \right] \\
 &= \left[\frac{1}{1 + e^{-z}} \right] \left[1 - \frac{1}{1 + e^{-z}} \right] \\
 &= s(z)(1 - s(z)) \\
 &= y(1 - y)
 \end{aligned}$$

En suivant l'équation du transparent précédent :

$$\begin{aligned}
 \nabla w_i &= -\eta \delta x_i^m \\
 &= -\eta (y - y^m) \frac{\partial s(z)}{\partial z} x_i^m \\
 &= -\eta (y - y^m) y(1 - y) x_i^m \\
 &= \eta (y - y^m) y(1 - y) x_i^m \\
 &= \eta y(1 - y) (y - y^m) x_i^m
 \end{aligned}$$

$$\begin{aligned}
 \nabla_w y(x^m, w) &= \left[\frac{\partial y}{\partial w_1}, \frac{\partial y}{\partial w_2}, \dots, \frac{\partial y}{\partial w_n} \right]^t \\
 z &= \sum_{i=0}^n w_i x_i \\
 y &= s(z) = \frac{1}{1 + e^{-z}}
 \end{aligned}$$



$$E = \sum_i \frac{1}{2} (y(x^m, w) - y^{i*})^2$$

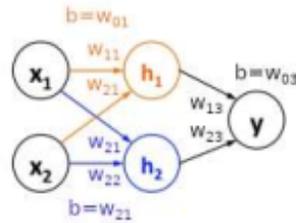
$$y = s \left(\underbrace{w_{13}s(\underbrace{w_{11}x_1 + w_{23}x_2 + w_{01}}_{z_1}) + w_{23}s(\underbrace{w_{12}x_1 + w_{22}x_2 + w_{02}}_{z_2}) + w_{03}}_{z_3} \right)$$

$$\frac{\partial E}{\partial w_j} = (y - y^{i*}) \frac{\partial y}{\partial w_j}$$

$$\frac{\partial y}{\partial w_{13}} = \frac{\partial y}{\partial z_3} \frac{\partial z_3}{\partial w_{13}} = \frac{\partial y}{\partial z_3} y_1 \tag{3.11}$$

$$\frac{\partial y}{\partial w_{11}} = \frac{\partial y}{\partial z_3} \frac{\partial z_3}{\partial w_{11}} = \frac{\partial y}{\partial z_3} \left(\frac{\partial y_1}{\partial z_1} \frac{\partial z_1}{\partial w_{11}} \right) = \frac{\partial y}{\partial z_3} \left(w_{13} \frac{\partial y_1}{\partial z_1} x_1 \right)$$

Exemple 3.8.1 Pour étudier la relation entre une maladie spécifique et le stress, la dépression, l'anxiété, laTerror et la moyenne de ces mesures, pour 40 données mas-



culins et féminins, les résultats ont été rassemblés dans le tableau suivant :

ID	stress	Anxiety	Terror	depression	Meen	Diagnosis
1	4	3	7	4	4,5	1
2	2	5	5	4	4	0
3	4	3	5	5	4,25	1
4	5	5	5	3	4,5	1
5	3	5	5	5	4,5	0
6	3	7	7	3	5	1
7	4	6	7	4	5,25	0
8	3	6	5	6	5	0
9	5	5	3	5	4,5	0
10	7	6	5	7	6,25	0
11	3	4	5	2	3,5	1
12	3	3	3	3	3	1
13	4	3	3	2	3	1
14	5	3	2	2	3	1
15	1	5	5	4	3,75	1
16	6	4	1	2	3,25	0
17	4	3	3	5	3,75	0
18	6	3	3	3	3,75	0
19	6	3	3	3	3,75	0
20	6	3	3	3	3,75	0
21	6	3	3	3	3,75	1
22	5	4	4	3	4	0
23	4	4	2	5	3,75	0
24	3	6	7	4	5	1
25	4	5	4	4	4,25	0
26	6	5	6	4	5,25	0
27	5	6	4	5	5	0
28	6	7	6	4	5,75	0
29	6	7	4	4	5,25	0
30	7	7	6	7	6,75	0
31	3	1	6	2	3	0
32	3	2	2	1	2	0
33	1	3	3	3	2,5	0
34	1	4	4	3	3	0
35	3	3	3	2	2,75	1
36	5	2	2	2	2,75	1
37	3	3	3	4	3,25	1
38	4	4	4	2	3,5	1
39	5	3	4	2	3,5	1
40	2	5	5	4	4	0

Nous supposons que le diagnostic de la maladie est la variable dépendante, le stress, l'anxiété, la terreur et la dépression, et la moyenne de ces mesures sont les variables indépendantes.

Nous obtenons les résultats suivants pour la régression linéaire multiple Ce tableau contient certaines des mesures qui ont été calculées pour le modèle estimé :

- La valeur du coefficient de corrélation $R = 0.476$, ce qui est très élevé.

Récapitulatif des modèles^b

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Modifier les statistiques				
					Variation de R-deux	Variation de F	ddl1	ddl2	Sig. Variation de F
1	,378 ^a	,143	,120	,46530	,143	6,342	1	38	,016

- Le carré du coefficient de corrélation est utilisé pour désigner l'étendue des données utilisées à partir des variables indépendantes

En estimant la variable dépendante, on remarque que le modèle estimé exprime (les deux variables indépendantes ensemble). 22% des données et une augmentation de la valeur de cette mesure expliquent que le modèle proposé est approprié.

- Définit le carré du coefficient de corrélation modifié et est utilisé dans le même but la précédente .
- Réglez l'erreur d'estimation, qui est ici de 0.46743, moins elle est significative sur l'erreur inférieure du modèle

Nous obtenons les résultats suivants pour le réseau de neurones artificiels

Statistiques

Diagnosis

N	Valide	40
	Manquant	0

Diagnosis

	Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide ,00	24	60,0	60,0	60,0
1,00	16	40,0	40,0	100,0
Total	40	100,0	100,0	

Classification

Echantillon	Observations	Prévisions		
		,00	1,00	Pourcentage correct
Apprentissage	,00	15	1	93,8%
	1,00	14	0	0,0%
	Pourcentage global	96,7%	3,3%	50,0%
Test	,00	6	2	75,0%
	1,00	2	0	0,0%
	Pourcentage global	80,0%	20,0%	60,0%

- Le premier tableau apparaît intitulé : diagnostic. nous avons 24 cas au moment du diagnostic et 16 cas dans le diagnostic 1
- Le deuxième tableau est intitulé : classification la prédiction a montré 37 cas dans

le diagnostic 0. Cela signifie qu'il y a 13 états d'erreur, Et que la prédiction a montré 3 cas dans le diagnostic 1. Pour comparer entre la méthode de régression linéaire multiple et la méthode des réseaux de neurones artificiels, nous utilisons l'erreur quadratique moyenne

$$MSE = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}. \quad (3.12)$$

cas appliqué	<i>MSE</i>
la régression linéaire multiple(RG)	1,162
Artificial Neural Networks(NN)	0.325

valeurs d'erreur quadratique moyenne standard pour le cas appliqué pour 40 échantillons

Conclusion

Grâce aux résultats des expériences de simulation présentées dans le tableau 1 pour 40 données, la méthode du réseau de neurone artificielle pour l'estimation de la régression multiple linéaire, basée sur le critère d'erreur quadratique moyenne est plus efficace que la méthode des moindres carrés ordinaires dans l'analyse de régression.

Bibliographie

- [1] ARNAUD-R & NICOLAS-V, « Econométrie, théorie et application », NATHAN, Paris, 1998.
- [2] Azaïs, J.-M., Bardet, J.-M. (2006). Le modèle linéaire par l'exemple. Régression analyse de la variance et plans d'expérience, Dunod, Paris.
- [3] Cornillon, P.-A., Matzner-LøBER, E. Régression avec R, Springer.
- [4] Greene, W, économétrie, 5e édition , New York university, Pearson Education France ,2005.
- [5] Guillaume Chevillon, « Econometrie » ,OFCE & Univ of Oxford, HEC 2005
- [6] Laudia Aroujo et al , « Econométrie : Repères- Cours – Applications », Edition Bréal ,2004.
- [7] Rzempoluck, E. J., (1998), “Neural Network Data Analysis Using SimulnetTM” , Springer.
- [8] Tham, M. T., (1996-1999), “Dynamic Models For Control Design” , Dep. Of Chemical and Process Engineering, University of Newcastle Upon Tyne, Newcastle Upon Tyne, U.K.

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$\mathbb{E}[X]$	Espérance mathématique ou moyenne du v.a X .
cov	Fonction de Covariance.
var	la variance mathématique
$\ \cdot\ $	est la norme euclidienne de \mathbb{R}^n
∇	gradient
$\sigma(\hat{\beta})$	désigne la valeur de l'écart-type du paramètre
Σ	la somme
MSE	l'erreur quadratique moyenne
$\mathbb{E}[\hat{\beta}]$	est sans biais.
$\hat{\beta}$	l'estimateur vecteur β
Π_X	la matrice de projection orthogonale
SCT	somme des carrés des écarts totale
SCE	somme des carrés des écarts expliquée
SCR	somme des carrés des écarts résiduelle
W	sont les poids du modèle,
f	est la fonction d'activation

Résumé

L'objectif de Mon mémoire est. d'étudier la régression multiple sous les trois aspects: théorique, pratique et bibliographique en s'appuyant sur les deux modèles neuronal et probabiliste. Du côté pratique en faisant apparaître la comparaison entre les deux méthodes d'estimation : les moindres carrés ordinaire et les réseaux de neurones

Mots clés: régression multiple, modèle de régression, réseau neuronal, descente de gradient.

Summary

The objective of My dissertation is. to study multiple regression under three aspects: theoretical, practical and bibliographical, relying on the two neural and probabilistic models. On the practical side by showing the comparison between the two estimation methods: ordinary least squares and neuron réseaux.

Key Word: multiple régression, régression model, neural network, gradient descente.

ملخص

الهدف من أطروحتي هو دراسة الانحدار المتعدد في جوانبه الثلاثة: النظرية والعملية والبيولوجرافية، وذلك بالاعتماد على النموذجين العصبي والاحتمالي. وعلى الجانب العملي من خلال إظهار المقارنة بين طريقتي التقدير: المربعات الصغرى العادية وخلايا العصبونات

الكلمات المفتاحية: الانحدار المتعدد، نموذج الانحدار، الشبكة العصبية، نزول متدرج