2.5 Analyse de différentes approches possibles sur un exemple

2.5.1 Approche probabiliste

On considère le problème de reconnaissance des chiffres à partir d'une image avec un bruit additif blanc gaussien.

Dans ce problème, on a $C = 10$ classes et pour une image $\boldsymbol{x}$ appartenant à la classe c, on a

$$\boldsymbol{x} = \boldsymbol{\mu}_c + \boldsymbol{\epsilon}$$

où $\boldsymbol{\mu}_c$ est l'image du chiffre de la classe c et $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{Id}\right)$.

Quelle est la loi suivie par $\boldsymbol{x}$?

Remarque : pour simplifier, on représente $\boldsymbol{x}$ et $\boldsymbol{\mu}$ par des vecteurs (même si ce sont des images).

A quelles conditions peut-on mettre en œuvre le discriminateur ci-dessous pour minimiser la probabilité d'erreur ?

$$\widehat{d} = \arg \max_{c=1,\ldots,10} \left(P(x|\omega_c)\right)$$

où

$$P(x|\omega_c) = \frac{1}{(2\pi)^{N/2}\,\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\mu}_c\|^2\right]$$

Or

$$\|\boldsymbol{x} - \boldsymbol{\mu}_c\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{\mu}_c\|^2 - 2\boldsymbol{\mu}_c^T\boldsymbol{x}$$

Donc

$$\widehat{d} = \arg\max_c \left(2\boldsymbol{\mu}_c^T\boldsymbol{x} - \|\boldsymbol{\mu}_c\|^2\right)$$

Qu'est-ce qu'on reconnaît dans l'équation ci-dessus ?

Quelle est la limite de cette approche ?

Parfois, il est nécessaire de rajouter des paramètres inconnus.

Par exemple, **lors du prochain TP, on considère le cas**

$$\boldsymbol{x} = A\,\boldsymbol{\mu}_c + \boldsymbol{\epsilon}$$

où $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d\right)$ et où A est une amplitude inconnue. La log-vraisemblance est

$$\log P = cte - \frac{1}{2\sigma^2}\|\boldsymbol{x} - A\boldsymbol{\mu}_c\|^2$$

et ainsi il faut minimiser la fonction

$$g = \|\boldsymbol{x} - A\boldsymbol{\mu}_c\|^2 = \|\boldsymbol{x}\|^2 + |A|^2\|\boldsymbol{\mu}_c\|^2 - 2A\boldsymbol{\mu}_c^T\boldsymbol{x}$$

Si on dérive par rapport à A

$$\frac{\partial g}{\partial A} = 2A\|\boldsymbol{\mu}_c\|^2 - 2\boldsymbol{\mu}_c^T\boldsymbol{x}$$

Donc, l'estimation de A est donnée par

$$\widehat{A} = \frac{\boldsymbol{\mu}_c^T\boldsymbol{x}}{\|\boldsymbol{\mu}_c\|^2}$$

On injecte cet estimateur dans g, on obtient une nouvelle fonction à optimiser :

$$\widetilde{g} = \|\boldsymbol{x}\|^2 + \frac{|\boldsymbol{\mu}_c^T\boldsymbol{x}|^2}{\|\boldsymbol{\mu}_c\|^4}\|\boldsymbol{\mu}_c\|^2 - 2\frac{\boldsymbol{\mu}_c^T\boldsymbol{x}}{\|\boldsymbol{\mu}_c\|^2}\boldsymbol{\mu}_c^T\boldsymbol{x}$$

et ainsi

$$\widetilde{g} = \|\boldsymbol{x}\|^2 + \frac{|\boldsymbol{\mu}_c^T\boldsymbol{x}|^2}{\|\boldsymbol{\mu}_c\|^2} - 2\frac{|\boldsymbol{\mu}_c^T\boldsymbol{x}|^2}{\|\boldsymbol{\mu}_c\|^2} = \|\boldsymbol{x}\|^2 - \frac{|\boldsymbol{\mu}_c^T\boldsymbol{x}|^2}{\|\boldsymbol{\mu}_c\|^2}$$

Ainsi, on obtient

$$\widehat{d} = \arg\max_c \left(\frac{|\boldsymbol{\mu}_c^T \boldsymbol{x}|^2}{\|\boldsymbol{\mu}_c\|^2} \right)$$

Qu'est-ce qu'on reconnaît ?

Quels paramètres inconnus supplémentaires est-il possible d'ajouter à la modélisation du problème ?

Avec cette approche est-il possible de résoudre le cas de chiffres manuscrits (pris en photo) ?

Exemple de solution pour des chiffres manuscrits.

Première étape. On suppose pouvoir prendre en photo le chiffre (avec le bon éclairage, le bon cadrage, la bonne orientation, le bon fond, ...).

Deuxième étape. On extrait le contour de l'image du chiffre, c'est à dire on estime la fonction

$$\begin{array}{ccc} [0,\,1] & \mapsto & \mathbb{C} \\ s & \mapsto & z(s) \end{array}$$

Comme le contour d'une forme est fermé, cette fonction est continue et 1-périodique. Ainsi, au lieu de manipuler une image avec N pixels, le chiffre est décrit par son contour, une fonction périodique.

Troisième étape. Pour simplifier la représentation de ce contour, on le décompose en série de Fourier. Ainsi au lieu d'avoir un contour, on a une représentation approximative décrite par un nombre fini de coefficients de la série de Fourier.

Quatrième étape. On combine des coefficients de la série de Fourier pour définir les "bons" paramètres pour réussir la discrimination des chiffres. Cette étude, non traitée ici faute de temps, est liée à la recherche des paramètres invariants du problème.

Cinquième étape. On effectue l'apprentissage avec ces "bons" paramètres.

Plus généralement, l'idée est d'extraire de la donnée initiale $\boldsymbol{x}$ des caractéristiques "utiles" $\boldsymbol{y}$

$$\begin{array}{ccc} \mathbb{R}^N & \mapsto & \mathbb{R}^d \\ \boldsymbol{x} & \mapsto & \boldsymbol{y} = \Phi(\boldsymbol{x}) \end{array}$$

avec $d << N$.

Quel est l'intérêt d'avoir $d << N$?

Est-ce qu'on est encore dans une approche probabiliste ?

2.5.2 Réseau de neurones avec une couche cachée

Comment un réseau de neurones permet de résoudre le problème considéré (reconnaissance des chiffres) ?

Les notations d'un réseau de neurones avec une couche cachée sont décrites Figure 11.
Le vecteur $\boldsymbol{w}_j$ (pour $j \in \{1, ..., N_c\}$) relie la couche d'entrée $\boldsymbol{x}$ au neurone j de la couche cachée

$$\nu_j = \sum_{i=0}^{N} x_i w_{ji} = \boldsymbol{w}_j^T \boldsymbol{x}$$

Le vecteur $\boldsymbol{w}_k$ (pour $k \in \{1, ..., N_s\}$) relie la sortie de la couche cachée $\boldsymbol{y}$ à l'entrée k de la couche de sortie

$$\eta_k = \sum_{j=0}^{N_c} y_j w_{kj} = \boldsymbol{w}_k^T \boldsymbol{y}$$

Entre l'entrée et la sortie d'un neurone, on a une fonction d'activation. Dans le TP 3, on choisit

$$f(z) = \frac{1 - e^{-2z}}{1 + e^{-2z}}$$

Ainsi, à la sortie on a

$$z_k = f(\eta_k) = f\left(\boldsymbol{w}_k^T \boldsymbol{y}\right) = f\left(\sum_{j=1}^{N_c} w_{kj} f\left(\sum_{i=1}^{N} w_{ji} x_i + w_{j0}\right) + w_{k0}\right)$$

Pour un problème de discrimination à plusieurs classes, on peut apprendre à un réseau de neurones à se rapprocher du cas où

$$z_k = \begin{cases} 1 \text{ si la classe de } \boldsymbol{x} \text{ est k} \\ \text{-1 sinon} \end{cases}$$

Dans ce cas, on définit alors comme discriminateur

$$\widehat{d} = \arg \max_{k=1,\ldots,N_s} z_k$$

Remarque. En terme d'estimation, l'apprentissage d'un réseau de neurones consiste donc à estimer les deux matrices $\boldsymbol{w}_j$ et $\boldsymbol{w}_k$.

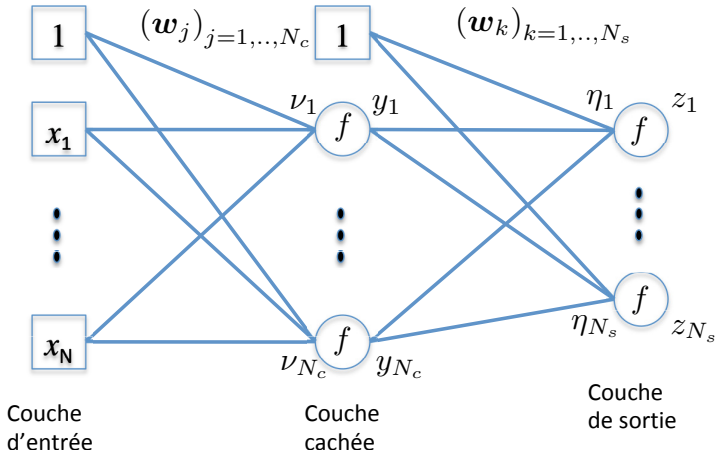


FIGURE 11 – Structure d'un réseau de neurones avec une couche cachée. Sur la couche d'entrée un vecteur de caractéristiques de taille N plus 1 offset (noté x_0), sur la couche cachée N_c neurones plus 1 offset et à la sortie N_s neurones.

Pourquoi utiliser un réseau de neurones avec une couche cachée peut être une bonne idée ?

Un peu d'histoire
Vers 1950, on se pose la question : comment résoudre le problème décrit figure 12 ?

Est-il possible d'arriver à de bonnes performances avec un discriminateur linéaire ?

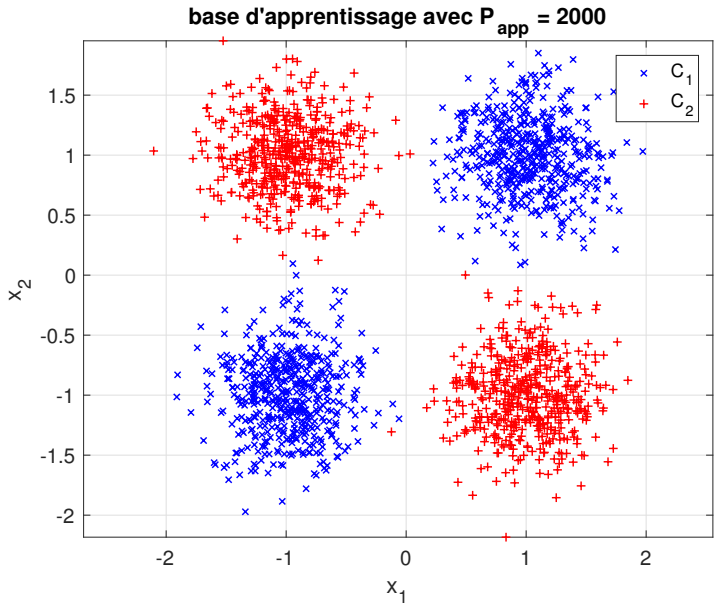


FIGURE 12 – Exemple du ou exclusif (XOR).

En utilisant le réseau de neurones suivant $N_c = 2$ et $N_s = 1$ et

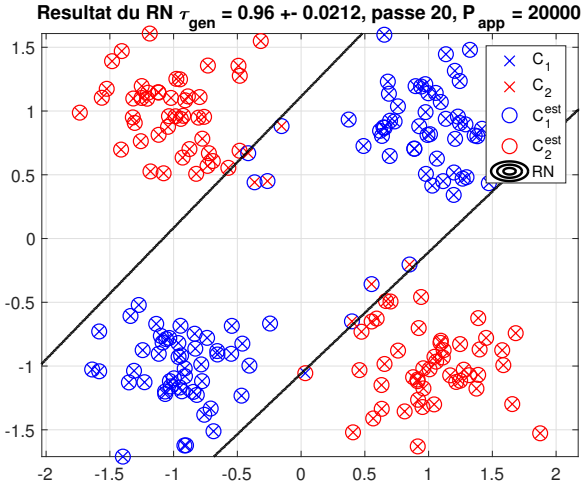
$$\begin{array}{lll} \boldsymbol{w}_{j=1} = (-2,\,-2,\,2) & \mapsto & \nu_1 = 2(x_2 - x_1 - 1) \\ \boldsymbol{w}_{j=2} = (-2,\,2,\,-2) & \mapsto & \nu_2 = 2(-x_2 + x_1 - 1) \\ \boldsymbol{w}_{k=1} = (3,\,3,\,3) & \mapsto & \eta = 3(y_1 + y_2) + 3 \end{array}$$

Une analyse simple (quoiqu'un peu fastidieuse) montre qu'on obtient

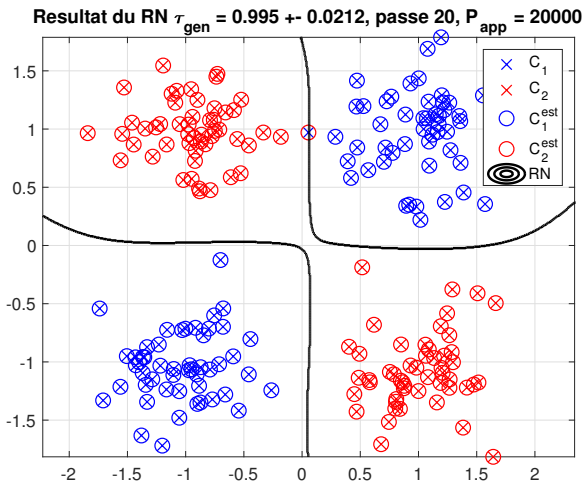
	ν_1	$y_1 = f(\nu_1)$	ν_2	$y_2 = f(\nu_2)$	η	$z = f(\eta)$
$x_1 \simeq x_2 \simeq 1$	$\nu_1 \simeq -2$	$y_1 \simeq -1$	$\nu_2 \simeq -2$	$y_2 \simeq -1$	$\eta \simeq -3$	$z \simeq -1$
$x_1 \simeq x_2 \simeq -1$	$\nu_1 \simeq -2$	$y_1 \simeq -1$	$\nu_2 \simeq -2$	$y_2 \simeq -1$	$\eta \simeq -3$	$z \simeq -1$
$x_1 \simeq -x_2 \simeq 1$	$\nu_1 \simeq -6$	$y_1 \simeq -1$	$\nu_2 \simeq 2$	$y_2 \simeq 1$	$\eta \simeq 3$	$z \simeq 1$
$x_1 \simeq -x_2 \simeq -1$	$\nu_1 \simeq 2$	$y_1 \simeq 1$	$\nu_2 \simeq -6$	$y_2 \simeq -1$	$\eta \simeq 3$	$z \simeq 1$

Comment voit-on que ce réseau de neurones avec une couche cachée permet effectivement de séparer les deux classes ?

La surface discriminante obtenue avec 2 neurones sur la couche cachée est représentée 13(a) et celle obtenue



(a) 2 neurones sur la couche cachée



(b) 6 neurones sur la couche cachée

FIGURE 13 – Les surfaces discriminantes obtenues avec un réseau de neurones pour une base d'apprentissage de taille $P_{app} = 2\,10^4$ sont tracées en trait noir.

avec 6 neurones sur la couche cachée est représentée 13(b).

Qu'observez-vous ?

Quels sont les problèmes ainsi rencontrés pour le réseau de neurones ?

Remarque. Grâce à d'une part l'augmentation des puissances de calculs, et d'autre part à la disponibilité de bases d'apprentissage de grande taille, des progrès importants ont été obtenus ces dix dernières années concernant cette étape d'apprentissage. Parmi les observations effectuées, il semble qu'une bonne stratégie est d'avoir plusieurs couches cachées.

En termes mathématiques, un réseau de neurones est simplement un mapping

$$\begin{array}{ccc} \mathbb{R}^N & \mapsto & \mathbb{R}^{N_s} \\ \boldsymbol{x} & \mapsto & \boldsymbol{z} = f(\boldsymbol{x}) \end{array}$$

Avec plusieurs couches, tout se passe comme si on avait plusieurs mappings

$$\begin{array}{ccccccc} \mathbb{R}^N & \mapsto & \mathbb{R}^{N_c^{(1)}} & \ldots & \mapsto & \mathbb{R}^{N_c^{(K)}} & \mapsto & \mathbb{R}^{N_s} \\ \boldsymbol{x} & \mapsto & \boldsymbol{y}_1 & \ldots & \mapsto & \boldsymbol{y}_K & \mapsto & \boldsymbol{z} = f(\boldsymbol{x}) \end{array}$$

Un point particulièrement intéressant est qu'il semble que, pour certaines applications, cette approche semble contribuer à **identifier** les caractéristiques utiles à discriminer.

Revenons sur notre problème de reconnaissance de chiffres.
Sachant que la solution optimale est

$$\widehat{d} = \arg \max_{c=1,..,10} \left(2\boldsymbol{\mu}_c^T \boldsymbol{x} - \|\boldsymbol{\mu}_c\|^2 \right)$$

qu'avec un réseau de neurones

$$\widehat{d} = \arg \max_{k=1,\ldots,N_s} z_k$$

et que

$$z_k = f(\eta_k) \quad \eta_k = \boldsymbol{w}_k^T \boldsymbol{y} \quad y_j = f(\nu_j) \quad \nu_j = \boldsymbol{w}_j^T \boldsymbol{x}$$

Comment fixer le nombre de neurones dans la couche cachée ?

Analysons à présent le choix du critère pour réaliser l'apprentissage du réseau de neurones.

On rappelle que le discriminateur de la pseudo-inverse a été obtenu en choisissant comme critère

$$J = \sum_{p=1}^{P_{app}} \|d^{(p)} - \boldsymbol{w}^T \boldsymbol{x}^{(p)}\|^2$$

où $d^{(p)} = \pm 1$ suivant la classe de $\boldsymbol{x}^{(p)}$.

Au vu des deux exemples ci-dessous, **le critère ci-dessus est-il adapté ?**
Si $d = 1$ et que $\boldsymbol{w}^T \boldsymbol{x} = -1$, le critère $J = 4$ et $\widehat{d} \neq d$.
Si $d = 1$ et que $\boldsymbol{w}^T \boldsymbol{x} = 3$, le critère $J = 4$ et $\widehat{d} = d$.

Le critère ci-dessous est-il adapté ?

$$J = \sum_{p=1}^{P_{app}} \|d^{(p)} - \text{sign}\left(\boldsymbol{w}^T \boldsymbol{x}^{(p)}\right)\|^2$$

A quoi peut servir de remplacer la fonction signe par la fonction ci-dessous ?

$$f(z) = \frac{1 - e^{-2z}}{1 + e^{-2z}}$$

Où a-t-on vu cette fonction précédemment ?

Soit $\boldsymbol{\eta} = (\eta_1, ..., \eta_{N_s})$ le vecteur obtenu à l'entrée de la couche de sortie du réseau de neurones. Pour chaque vecteur $\boldsymbol{x}^{(p)}$ de la base d'apprentissage, on a un vecteur

$$\boldsymbol{\eta}^{(p)} = \sum_{j=1}^{N_c} w_{kj} f\left(\boldsymbol{w}_j^T \boldsymbol{x}^{(p)}\right)$$

Dans le cas à deux classes, le critère utilisé pour effectuer l'apprentissage du réseau de neurones est

$$J = \sum_{p=1}^{P_{app}} \| d^{(p)} - f\left(\boldsymbol{\eta}^{(p)}\right) \|^2$$

Dans le cas à $N_s > 2$ classes, le critère utilisé pour effectuer l'apprentissage du réseau de neurones est

$$J = \sum_{p=1}^{P_{app}} \| \boldsymbol{t}^{(p)} - f\left(\boldsymbol{\eta}^{(p)}\right) \|^2$$

où $\boldsymbol{t}^{(p)}$ est le vecteur cible (en anglais "target") qui vaut -1 sur toutes les composantes sauf sur la composante qui correspond à sa classe où il vaut 1.

Quels sont les problèmes potentiels lors de la mise en œuvre d'un réseau de neurones ?

Lors de la séance de travaux pratique n°3, on analysera les performances d'un réseau de neurones pour la reconnaissance des chiffres.

Lors de la séance de travaux pratique n°4, on étudiera l'applicabilité de la théorie du risque Bayésien.