Introduct° à la Reconnaissance de formes Statistiques

Schéma général

$\bar{y}$ (observat°) ⟶ Extraction caractéristiques ⟶ $\bar{x}$ ⟶ classifieur ⟶ $\left\{ \begin{array}{l} \text{Décision} \\ \text{classificat°} \end{array} \right.$

↳ Dans ce cours

→ On a $\bar{x} \in \mathbb{R}^n$, $\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$, $\exists$ M classes $w_1, w_2, \ldots, w_M$

$\boxed{\text{Objectif}}$: Déterminer dans quelle classe ranger $\bar{x}$

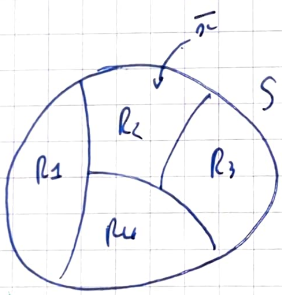
c.a.d : À partir de $\bar{x}$ ⟹ Quelle classe lui attribuer

Méthodes déterministes

$\bar{x} \in S \subset \mathbb{R}^n$

$\bigcup_{m=1}^{M} R_m = S$ (exhaustivité)

$R_i \cap R_j = \emptyset$ si $i \neq j$ (non-ambiguité)



$\boxed{\bar{x} \in R_m \Leftrightarrow \bar{x} \text{ est dans la classe } w_m}$

Voc : $M = 2$ ⟶ "Détection"
$M > 2$ | moins connu ⟶ "Discriminat°"
| moins inconnu ⟶ "classificat°"

I. Méthodes Linéaires de Discrimination

A/ Cas à M = 2 classes

→ Problématique : On a 2 bases d'exemples pour l'apprentissage

$\mathcal{B}_{A1} = \{ \bar{x}^l \mid \bar{x}^l \in w_1 \}$ $\qquad \mathcal{B}_A = \mathcal{B}_{A1} \cup \mathcal{B}_{A2}$

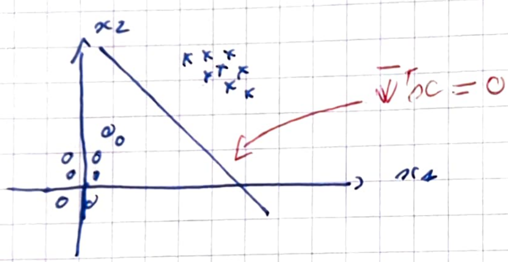
$\mathcal{B}_{A2} = \{ \bar{x}^l \mid \bar{x}^l \in w_2 \}$

à chaque $\bar{x}_l \in \mathcal{B}_A \longrightarrow d^l = \begin{cases} 1 \text{ si } \bar{x}^l \in w_1 \\ -1 \text{ ——— } w_2 \end{cases}$

Objectif de la DL : déterminer un vecteur $\bar{w} \in \mathbb{R}^N$ tq :

$$\bar{w}^T \bar{x} + w_0 \quad \begin{matrix} > 0 \text{ si } \bar{x} \in w_1 \\ < 0 \text{ si } \bar{x} \in w_2 \end{matrix}$$

$$\bar{w}^T \bar{x} = \sum_i w_i x_i = \text{Produit scalaire entre } \bar{w} \text{ et } \bar{x}$$



$\bar{w}^T \bar{x} + w_0 = 0$: Eq d'1 Hyperplan. Notat° : $\bar{w}' = \begin{pmatrix} w_0 \\ \vdots \\ w_N \end{pmatrix}$, $\bar{x}' = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_N \end{pmatrix}$

$\Rightarrow \bar{w}^T \bar{x} + w_0 = \bar{w}'^T \bar{x}'$ ⟶ écriture + simple

Dans la suite, on utilisera $\bar{w}'$ et $\bar{x}'$ sans noter les " ' " i.e $\{ \bar{w}, \bar{x}$

Méthodologie générale : Base d'exemple → B_A apprentissage (taille P_A) ; → B_G généralisat° (taille P_G)

B/ Méthode de Hebb

$$\bar{w}_{Hebb} = \sum_{l=1}^{P_A} d^l \bar{x}^l$$

(d^l : +1 / −1)

Caractérisat° des performances :

Tx d'apprentissage : $\tau_A = \frac{1}{P_A} \sum_{l=1}^{P_A} m_l$

$m_l = \begin{cases} 1 \text{ si } \operatorname{sign}(\bar{w}^T \bar{x}^l) = \operatorname{sign}(d^l) \text{ : pas d'erreur} \\ 0 \text{ sinon : Erreur} \end{cases}$

Tx généralisat° : $\tau_G = \frac{1}{P_G} \sum_{l=1}^{P_G} m_l$

C/ Méthode de la Pseudo - Inverse

- Critère plus rigoureux : erreur quadratique

$$J = \sum_{\ell=1}^{p_1} \left[\bar{w}^T \bar{x}^{\ell} - d_\ell \right]^2 = \| \bar{d} - \bar{\bar{X}} \bar{w} \|^2$$

avec $\left\{ \begin{array}{l} d = \begin{pmatrix} d_1 \\ \vdots \\ d_{p_1} \end{pmatrix} \\ \bar{\bar{X}} = \begin{pmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_N^1 \\ \vdots & & & & \vdots \\ 1 & x_1^{p_1} & x_2^{p_1} & \cdots & x_N^{p_1} \end{pmatrix} = \begin{pmatrix} x^{1T} \\ \vdots \\ x^{p_1 T} \end{pmatrix} \end{array} \right.$

<u>Objectif</u> : Trouver $\bar{w}$ qui minimise J

$$J = [\bar{d} - \bar{\bar{X}}\bar{w}][\bar{d} - \bar{\bar{X}}\bar{w}] = \underbrace{d^T d}_{\text{scalaire}} - d^T X w - w^T X^T d + w^T X^T X w$$

$$= a - 2 w^T \underbrace{X^T d}_{\bar{v}} + w^T \underbrace{X^T X}_{\bar{\bar{M}}} w$$

$$= a - 2 w^T v + w^T M w = a - 2 \sum_{i=1}^{N+1} w_i v_i + \sum_{i,j} w_i M_{ij} w_j$$

$$\frac{\partial J}{\partial w_m} = -2 v_m + \sum_j M_{mj} w_j + \sum_i M_{im} w_i \quad \text{or } M_{im} = M_{mi} \text{ car } M^T = M \; (M = X^T X)$$

$$= -2 v_m + 2 \sum_i M_{mi} w_i$$

d'où $\dfrac{\partial J}{\partial w_m} = 0 \iff \sum_i M_{mi} w_i = v_m \quad \forall m$

$$\iff \bar{\bar{M}} \bar{w} = \bar{v}$$

$$\iff \bar{\bar{X}}^T \bar{\bar{X}} \bar{w} = \bar{\bar{X}}^T \bar{d}$$

Finalement, on cherche $\bar{w}$ tq $\boxed{\bar{\bar{X}}^T \bar{\bar{X}} \bar{w} = \bar{\bar{X}}^T \bar{d}}$

<u>Solut°</u> : * Cas 1 : $X^T X$ est inversible. Alors $\bar{w} = \underbrace{(\bar{\bar{X}}^T \bar{\bar{X}})^{-1} \bar{\bar{X}}^T}_{\bar{\bar{X}}^{PI}} \bar{d}$ Pseudo inverse

* Cas 2 : $X^T X$ non inversible

$\bar{\bar{M}} = \bar{\bar{X}}^T \bar{\bar{X}}$ est symétrique $\Rightarrow$ diagonalisable

i.e $\exists U$ normale ($\bar{U}^T \bar{U} = \bar{U} \bar{U}^T = Id$) tq $\bar{\bar{M}} = U^T \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix} U$

Soit r le rang de M. Si $r < N+1$, M non inversible (rang = nb de vp non nul)

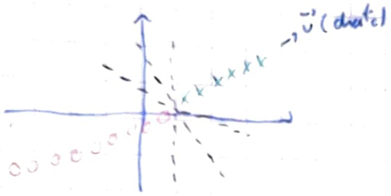
$$\bar{\bar{M}} = U^T \begin{pmatrix} \lambda_1 & & & & 0 \\ & \ddots & & & \\ & & \lambda_r & & \\ & & & 0 & \\ 0 & & & & 0 \end{pmatrix} U$$

<u>Pseudo - inverse de M</u> : $\bar{\bar{M}}^{PI} = U^T \begin{pmatrix} 1/\lambda_1 & & & & 0 \\ & \ddots & & & \\ & & 1/\lambda_r & & \\ & & & 0 & \\ 0 & & & & 0 \end{pmatrix} U$

On a bien $\bar{\bar{M}}^{PI} = M^{-1}$ si rang $r = N+1$

$\overline{w} = (X^T X)^{PI} X^T d$: généralisat° de la PI de $\overline{\overline{X}}$

Exemple : $N=2$



TP : Parfois quand $P_A \sim N$, il y a des λ_i très petit donc des coef $\frac{1}{\lambda_i}$ très grand.

Solut° → Augmenter taille base avec du bruit

D/ Limites de la PI

$$J = \sum_{l=1}^{P_A} (w^T x^l - d_l)^2$$

Il faudrait plutôt utiliser :

$$J = \sum_{l=1}^{P_A} |\, \text{sign}(w^T x^l) - d_l \,| \rightarrow \text{mais } \not\exists \text{ solut° analytique i.e } \not\exists \text{ gradient}$$

Ex 1 : • Un l donné, $d_l = 1$, $w^T x^l = -1$

$(w^T x^l - d_l)^2 = 4$ erreur discriminat°

• $d_l = 1$ et $w^T x^l = 3$ donc $(w^T x^l - d_l)^2 = 4$ Pas d'erreur

⟹ Le critère n'est pas le bon !

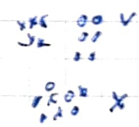
Critère du Perceptron : $\boxed{J_p = \sum_{l \in \Omega_w} - d_l w^T x^l}$, $\Omega_w = \{ l \text{ tq } d_l w^T x^l < 0 \}$

donc, $J_p \geq 0$ et $J_p = 0$ si aucune erreur

Méthode Gradient : $\overline{w}^{k+1} = \overline{w}^{k} - \alpha(k) \vec{\nabla} J_p$

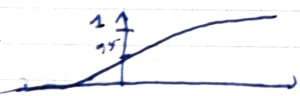
Algorithme du Perceptron → converge vers une solut° si B_A est linéairement séparable

Sinon : problèmes de convergence



Discriminat° logistique (Cox : 1960)

$g(x) = \frac{e^x}{1+e^x}$



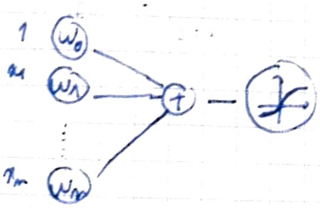
Nouveau critère :

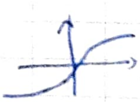
$$\boxed{J_{DL} = \sum_{l=1}^{P_A} [\alpha_l - g(\overline{w}^T \overline{x})]^2}$$

→ Fonct° J dérivable ⟹ Méthode de gradient pour trouver w

$\alpha_l = \frac{d_l + 1}{2} = \begin{cases} +1 & \text{si } x^l \in w_1 \\ 0 & \text{— } w_2 \end{cases}$

Autre écriture du m̂ type : $J_{RN} = \sum_{l=1}^{P_A} [d_l - \sigma(w^T x)]^2$ où $\sigma(x) = \frac{e^x - 1}{e^x + 1} = 2g(x) - 1$





Réseaux de neurones