

Cours: Topics in Machine Learning

Plan

- 1 Éléments sur les réseaux de neurones (NN) E.R.
- 2 Théorie de l'approximation pour les NN E.A.
- 3 Complexité des NN E.R.
- 4 Régression non paramétrique avec les NN IC
- 5 Generative Adversarial Network E.A.
- 6 Interpolation vs Surapprentissage C.B.
- 7 Apprentissage et confidentialité E.A.

Chapitre 1 : Complexité des réseaux de neurones

Plan

- ① Dimension VC d'une classe de réseaux de neurones
coeffcient d'éclatement, dimension de Vapnik - Chervonenkis (VC)
rôle en classification pour le contrôle de l'erreur stochastique
majoration d'une classe de NN à support fixé
- ② Entropie d'une classe de réseaux de neurones
nombre de recouvrements , entropie, rôle en régression non-paramétrique
majoration de l'entropie d'une classe de NN à opacité fixée

1 Dimension VC d'une classe de NN

1.1 Rappels sur la dimension VC

Définition: Soit \mathcal{H} un ensemble de fonctions de \mathbb{R}^d dans $\{-1, 1\}$ "classificateurs"

Le coefficient d'éclatement de \mathcal{H} de $m \geq 1$ points est donné par

$$S_{\mathcal{H}}(m) = \max_{x_1, \dots, x_m \in \mathbb{R}^d} \#\{(h(x_1), \dots, h(x_m)) \mid h \in \mathcal{H}\}$$

nombre maximum d'étiquetages de m points que la classe \mathcal{H} peut produire

$$\text{On a } \forall m \geq 1, \quad 1 \leq S_{\mathcal{H}}(m) \leq 2^m$$

Si $S_{\mathcal{H}}(m) = 2^m$, alors \mathcal{H} "éclate" un échantillon de taille m
↳ mesure de la taille de \mathcal{H}

Définition: la dimension VC de \mathcal{H} est donnée par

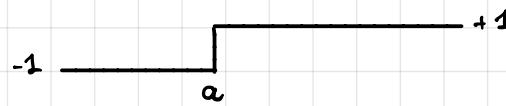
$$d_{VC}(\mathcal{H}) = \sup \{ m \geq 0 : S_{\mathcal{H}}(m) = 2^m \} \quad (S_{\mathcal{H}}(0) = 1)$$

la taille maximum d'un échantillon que \mathcal{H} peut éclater
le nombre maximum de points qui peuvent être arbitrairement
étiquetés par \mathcal{H}

Remarque: si $S_{\mathcal{H}}(m) = 2^m$ et $S_{\mathcal{H}}(m+1) < 2^{m+1}$, alors $d_{VC}(\mathcal{H}) = m$

Exemples:

* $\mathcal{H} = \{R_0\}$: $S_{\mathcal{H}}(m) = 1$ pour tout m , et $d_{VC}(\mathcal{H}) = 0$

* $\mathcal{H} = \{\text{sign}(.-a), a \in \mathbb{R}\}$ ($d=1$) 

$S_{\mathcal{H}}(1) = 2$, $S_{\mathcal{H}}(2) = 3$ (10 non!) donc $d_{VC}(\mathcal{H}) = 1$

* $\mathcal{H} = \{2^n \mathbf{1}_{[a \leq x \leq a+2^n]} : n \in \mathbb{N}, a \in \mathbb{R}\}$ 

$S_{\mathcal{H}}(1) = 2$, $S_{\mathcal{H}}(2) = 4$, $S_{\mathcal{H}}(3) < 2^3$, $d_{VC}(\mathcal{H}) = 2$ (111 et 010 non!)

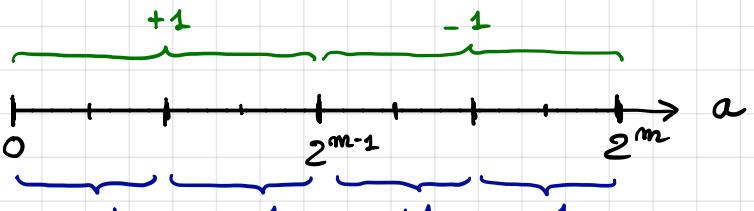
* Si $\alpha = \{ \text{sign}(\sin(\pi \alpha \cdot)), \alpha \in \mathbb{Q}_2 \}$ ($\alpha = \frac{p}{q}$) alors $d_{VC}(\alpha) = +\infty$
 $(x_1, \dots, x_m) = (2^{-k_i}, k=1 \dots m)$ peut être éclaté par α .

Choix de α ?

$$\sin(\pi \alpha 2^{k-1}) = \begin{cases} +1 \\ -1 \end{cases} \quad \text{pour} \quad \begin{cases} \alpha \in](2i)2^{k-1}, (2i+1)2^{k-1}[\\ \alpha \in](2i+1)2^{k-1}, (2i+2)2^{k-1}[\end{cases}, i \in \mathbb{Z}$$

on choisit le signe de $\sin(\pi \alpha x_m)$, $x_m = 2^{cm-1}$ librement

$\text{sign}(\pi \alpha x_m)$



$\text{sign}(\pi \alpha x_{m-1})$

$\text{sign}(\pi \alpha x_{m-2})$ \dots

et donc choix de α possible pour donner un vecteur de signes arbitraire

En particulier, $d_{VC}(\alpha)$ non lié au nombre de paramètres définissant α .

* Si $\mathcal{H} = \{x \in \mathbb{R}^d \mapsto \text{sign}(a^T x + b), a \in \mathbb{R}^d, b \in \mathbb{R}^d\}$

"classifieurs affines"

pour $d=1$, $\mathcal{H} \supset \{\text{sign}(c - x), c \in \mathbb{R}\} \cup \{\text{sign}(-c - x), c \in \mathbb{R}\}$



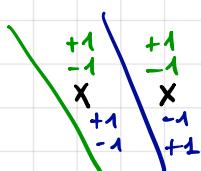
$$\text{donc } \mathcal{D}_{VC}(\mathcal{H}) = 2$$

pour $d=2$, on peut faire un dessin

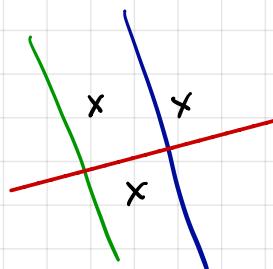
$$m=1$$



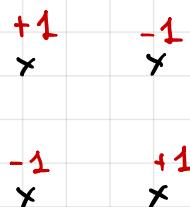
$$m=2$$



$$m=3$$



$$m=4$$



?

$$S_{\mathcal{H}}(1) = 2$$

$$S_{\mathcal{H}}(2) = 2^2$$

$$S_{\mathcal{H}}(3) = 2^3$$

$$S_{\mathcal{H}}(4) < 2^4$$

$$\text{donc } \mathcal{D}_{VC}(\mathcal{H}) = 3$$

Lemme : pour $d \geq 1$, $\mathcal{D}_{VC}(\mathcal{H}) = d+1$

Exercice 9.5.2 de [Giraud (2015)]

1.2) Rôle de la dimension VC en classification

(X_i, Y_i) , $1 \leq i \leq n$ copies iid de $(X, Y) \in \mathbb{R}^d \times \{-1, +1\}$

g^e ensemble de fonctions de $\mathbb{R}^d \rightarrow \{-1, +1\}$ classifiants

$L(R) = \mathbb{P}(R(X) \neq Y)$, $R \in g^e$ perte

$\hat{R}_{ge}^* \in \operatorname{argmin}_{R \in g^e} L(R)$ oracle dans g^e

$\hat{R}_n \in \operatorname{argmin}_{R \in g^e} \left\{ \sum_{i=1}^n \mathbb{I}\{R(X_i) \neq Y_i\} \right\}$ minimisateur du risque empirique

Théorème avec proba $\geq 1 - e^{-t}$ ($t > 0$)

$$L(\hat{R}_n) - L(\hat{R}_{ge}^*) \leq 4 \sqrt{\frac{2 \text{VC}(g^e) \log(2n+2)}{n}} + \sqrt{\frac{2t}{n}}$$

Contrôle de l'erreur stochastique

pas du biais

$$L(\hat{R}_{ge}^*) - \min_{R \text{ mesurable}} L(R)$$

Majoration de la dimension VC pour une classe de NN

Rappels:

- * un NN est $\Phi = (W, \rho)$ avec

$\left\{ \begin{array}{l} \text{dim entrée } d \\ \text{dim sortie } k \\ W = ((A_1, b_1), \dots, (A_L, b_L)) \text{ poids} \\ \rho: \mathbb{R}^d \rightarrow \mathbb{R}^k \text{ fct activation} \end{array} \right.$

profondeur de Φ est L

A_ℓ matrice réelle $N_\ell \times N_{\ell-1}$

N_ℓ = nombre de neurones (largeur) de la couche ℓ ; $N = \sum_{\ell=1}^L N_\ell$ largeur totale

sparsité de la couche ℓ est $\|A_\ell\|_0 + \|b_\ell\|_0$

support de la couche ℓ est les indices où ces poids sont $\neq 0$

- * la réalisation de Φ est donnée par la fonction

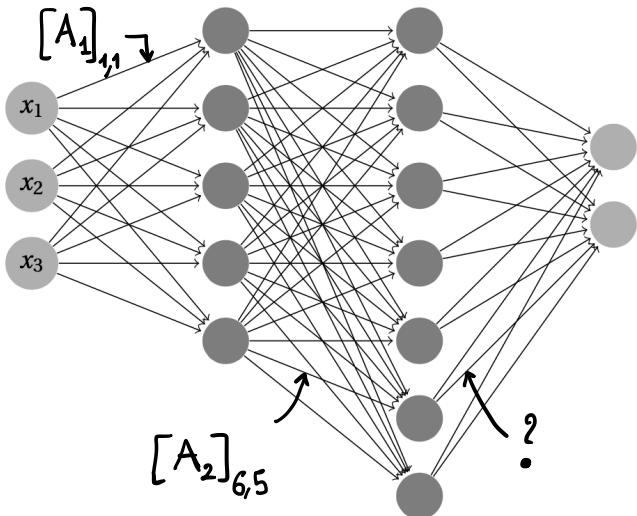
$$R(\Phi) = T_L \circ \rho \circ T_{L-1} \circ \dots \circ T_2 \circ \rho \circ T_1$$

où $T_\ell: x \in \mathbb{R}^{N_{\ell-1}} \mapsto A_\ell x + b_\ell \in \mathbb{R}^{N_\ell}$ ($N_0 = d, N_L = k$)

Représentation d'un NN

(en fait de la suite des A_i surtout)

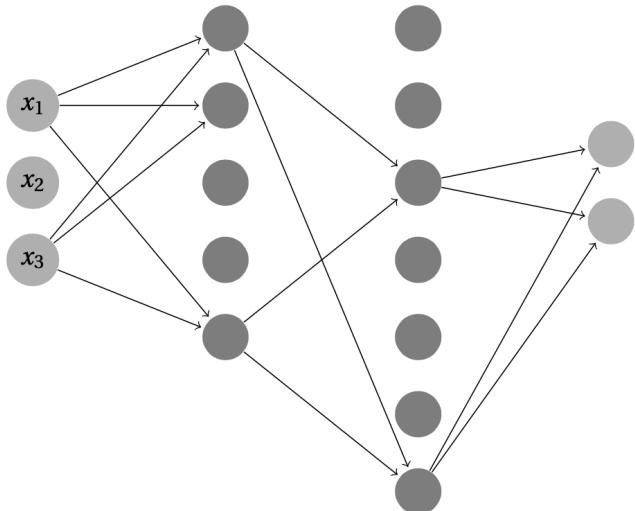
Exemple 1



$$\begin{cases} L = 3 \\ \delta = 3, k = 2, N_1 = 5, N_2 = 7, N_{\max} = 7 \end{cases}$$

entièrement connecté ($[A^e]_{ij} \neq 0$)
 $\forall e, i, j$

Exemple 2



idem mais sparse

$$\text{sparsité } \|\Phi\|_0 = 14$$

(si $b^e = 0 \quad 1 \leq e \leq L$)

Classification à l'aide d'une famille de NN

On considère des NN PreLU $\{\Phi_\omega, \omega \in \mathbb{R}^s\}$

avec dimension d'entrée d , de sortie N_L , de profondeur L

longueur de la couche l , $N_l \geq 1$ avec $N = \sum_{l=1}^L N_l \geq 3$

sparsité de la couche l , $\|A_l\|_0 + \|b_l\|_0 \leq S_l$ avec $S = \sum_{l=1}^L S_l \geq N$

support fixé correspondant aux $(S_l)_{1 \leq l \leq L}$

Tous les (A_l, b_l) se résument donc en un vecteur de poids $\omega \in \mathbb{R}^s$

Au final, on considère la famille de classifieurs

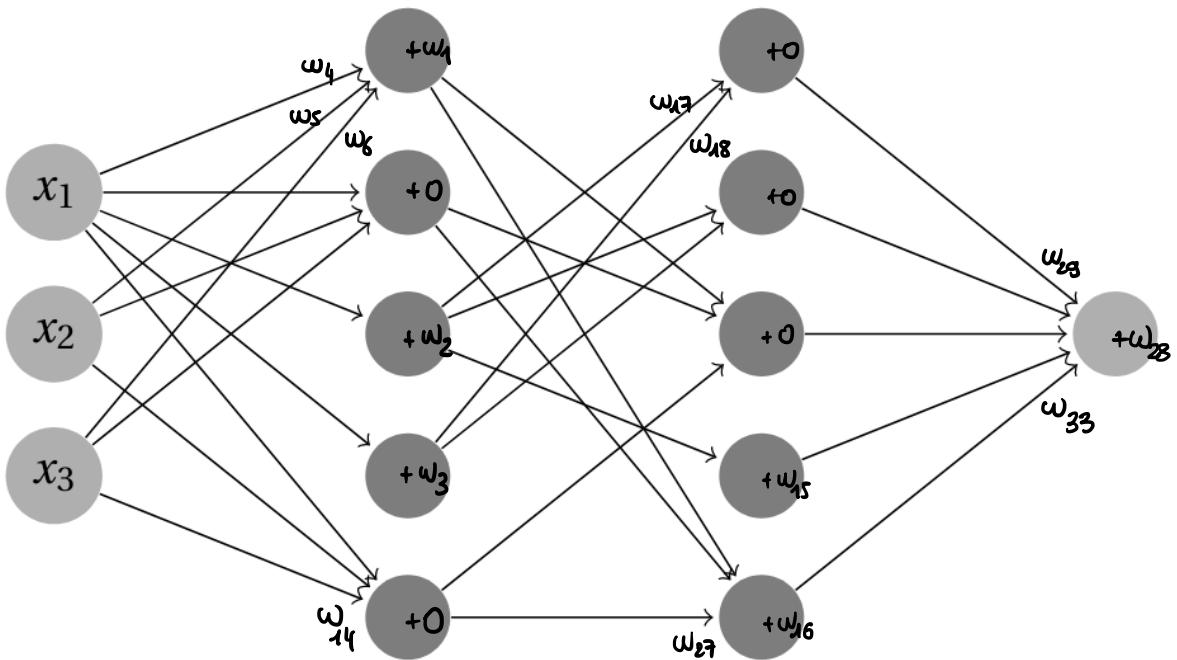
$$\mathcal{H} = \left\{ \text{sign}(\Phi(\Phi_\omega)), \omega \in \mathbb{R}^s \right\}$$

Illustration

$$S_1 = 14$$

$$S_2 = 13$$

$$S_3 = 6$$



$s = \text{nombre d'arêtes} + \text{dimension des poids constants} = 33$
 Au final, paramétrisé par $\omega \in \mathbb{R}^{33}$

nb de coefficients "achifs"

Borne sur l'ce dimension VC

$$\mathcal{H} = \left\{ \text{sign}(\Phi(\omega)), \omega \in \mathbb{R}^s \right\}$$

avec une topologie fixée (L, N, s) comme plus haut

Théorème

$$S_{\mathcal{H}}(m) \leq (4e m L)^{sL} \quad \text{pour } m \geq s$$

$$d_{VC}(\mathcal{H}) \leq 6sL \log(4eN)$$

Idée: $\omega \in \mathbb{R}^s \mapsto R(\Phi_\omega)$ est polynomiale par morceaux (avec s variables)

Remarques:

Preuve précise plus loin

* il existe une borne inférieure qui recouvre presque

$\exists c > 0$ tq $\forall s \geq cL, \forall L \geq c$, \exists RéLU NN de profondeur L et complexité s

avec $d_{VC}(\mathcal{H}) \geq sL \log(s/L)/c$ Théorème 3 [Bartlett et al (2019)]

* mesure précisément la complexité en fonction de s, L, N

Propriété: dans le cadre du théorème, pour tout $x_1, \dots, x_m \in \mathbb{R}^d$, il existe une partition $(P_i)_{1 \leq i \leq M}$ de \mathbb{R}^s de taille $M \geq 1$ ($P_i = \emptyset$ éventuellement) telle que :

* pour tout $1 \leq i, j \leq M$, $w \in P_i \mapsto R(\phi_w)(x_j)$

coincide avec un polynôme multivarié en s variables de degré $\leq L$

* $M \leq \frac{1}{\epsilon^L} \alpha \left(2emL N \rho / \bar{s}_\ell \right)^{\bar{s}_\ell}$, où $\bar{s}_\ell = \sum_{i=1}^\ell s_i$

Preuve: par récurrence le long des couches du NN

on construit une suite de \mathbb{R}^s -partitions $S_0 = \mathbb{R}^s, S_1, \dots, S_{L-1}$ imbouties

avec ① $\forall \ell \in \{1, \dots, L-1\}, \#S_\ell / \#S_{\ell+1} \leq 2 \left(2emL N \rho / \bar{s}_\ell \right)^{\bar{s}_\ell}$

② $\forall \ell \in \{1, \dots, L-1\}, \forall S \in S_\ell, \forall j \in \{1, \dots, m\}$,

la sortie du neurone de la ℓ ème couche prenant x_j en entrée coïncide avec un polynôme multivarié en $w \in S$ de degré $\leq \ell$

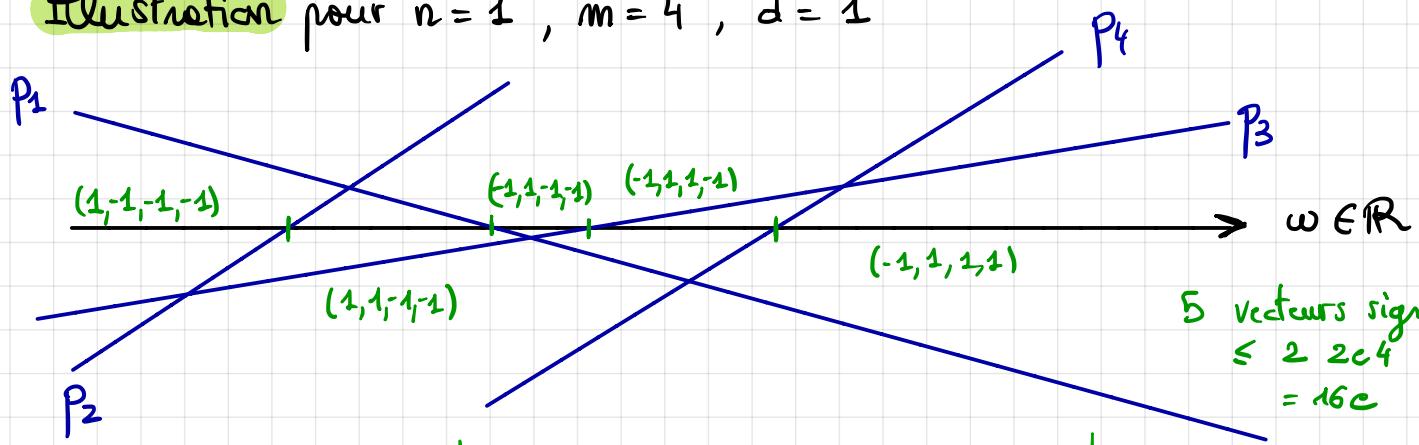
Lemma oechli: Soit $p_1 \dots p_m$ des polynômes multivariés
en $n \leq m$ variables
de degré (total) au plus $d \geq 1$

$$\text{Alors } \#\{(\text{sign}(p_1(\omega)), \dots, \text{sign}(p_m(\omega))) , \omega \in \mathbb{R}^d\} \leq 2 (2e^{md/n})^n$$

nb vecteurs signes générés par les valeurs de ces m polynômes

Preuve: cf Anthony et al. (1990) (Theorem 8.3)

Illustration pour $n = 1$, $m = 4$, $d = 1$



pour $d \geq 1$, au plus d racines, ce qui multiplie par d

$$5 \text{ vecteurs signes} \leq 2 \cdot 2^4 = 16$$

Discussion

* l'argument se généralise à d'autres fonctions

d'activation, polynomiale par morceaux | avec $p+1$ morceaux
de degré $r \geq 0$

$$r=0, \quad d_{VC}(\mathcal{H}) \lesssim s \log(p_s)$$

$$r \geq 1, \quad d_{VC}(\mathcal{H}) \lesssim L s \log(p_N) + L^2 s \log r$$

* Rôle de la profondeur

Ainsi : $L \nearrow \Rightarrow$ risque stochastique \nearrow ie variance \nearrow

Chapitre 1 & 2 : $L \nearrow \Rightarrow$ approx meilleure ie biais \searrow

Prendre un NN profond n'est pas toujours une bonne solution

Il s'agit de réaliser un bon compromis biais-variance cf chap 4

② Entrée d'une classe de réseaux de neurones

2.1 Rappels sur le nombre de recouvrements

Définition: Soit $(E, \|\cdot\|)$ un espace vectoriel normé

et $A \subset E$ un sous-ensemble quelconque

Pour $\delta \in (0, 1)$, le nombre de recouvrement de A est

$$N(\delta, A, \|\cdot\|) = \min \{ k \geq 1 : \exists e_1, \dots, e_k \in E \text{ tq } A \subset \bigcup_{i=1}^k B_{\frac{\delta}{2}}(e_i, \delta) \}$$

où $B_{\frac{\delta}{2}}(e_i, \delta) = \{ x \in E : \|x - e_i\| \leq \frac{\delta}{2} \}$ boule centrée en e_i de rayon $\frac{\delta}{2}$

L'entropie de A est $\log N(\delta, A, \|\cdot\|)$

nombre minimum de $\|\cdot\|$ -boules de rayon δ pour recouvrir A

Exemples

1) $(E, \|\cdot\|) = (\mathbb{R}, |\cdot|)$ et $A \subset [-c, c]$ pour $c > 0$

on pose $\mathbb{A} = [2c/\delta]$

et $(x_1, \dots, x_k) = (-c + \delta, -c + 2\delta, \dots, -c + k\delta)$

de sorte que $x_k = -c + k\delta \leq c$ et $x_{k+1} = -c + (k+1)\delta > c$

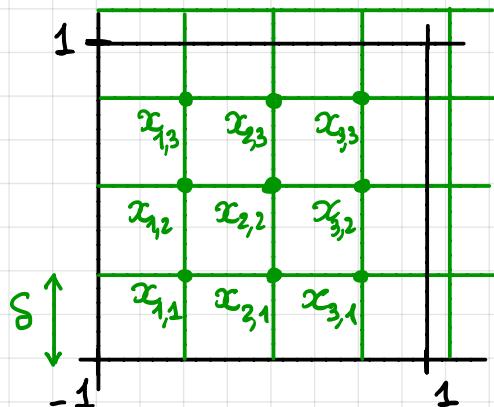
Ainsi $\mathcal{N}(S, A, 1 \cdot) \leq 2c/\delta$

2) $(E, \|\cdot\|) = (\mathbb{R}^d, \|\cdot\|_\infty)$ et $A \subset [-1, 1]^d$

On construit la grille plus fine
sur chaque axe, ce qui donne
les centres $(x_{i_1, \dots, i_d}, 1 \leq i_1, \dots, i_d \leq k)$

Ainsi

$\mathcal{N}(S, A, \|\cdot\|_\infty) \leq (2c/\delta)^d$



3) $(E, \|\cdot\|_1) = (\mathbb{R}^d, \|\cdot\|_\infty)$, $s \in \{1, \dots, d\}$ sparsité

$$A = \{x \in \mathbb{R}^d : \|x\|_0 \leq s, \|x\|_\infty \leq 1\}$$

Alors $A \subset \bigcup_{0 \leq r \leq s} \bigcup_{S \subset \{1, \dots, d\}} A_S$ sparse and bounded

$$\|x\|_0 = \sum_{i=1}^d \mathbb{1}_{\{x_i \neq 0\}}$$

$\|x\|_0 \leq s \Leftrightarrow$ au plus s coordonnées $\neq 0$

$$A_S = \{x \in \mathbb{R}^d : \forall i \notin S, x_i = 0, \forall i \in S, |x_i| \leq 1\}$$

D'après le cas 2) en dimension $|S|=r$ pour $R \leq (2/s)^r$

$$\exists y_1, \dots, y_R \in \mathbb{R}^S \text{ tq } [-1, 1]^S \subset \bigcup_{i=1}^R \{y \in \mathbb{R}^S : \|y - y_i\|_\infty \leq s\}$$

donc $\exists x_1, \dots, x_R \in \mathbb{R}^d$ tq $A_S \subset \bigcup_{i=1}^R \{x \in \mathbb{R}^d : \|x - x_i\|_\infty \leq s\}$

\uparrow idem t unei sur S et aussi sur S^c !

$(x_i)_{i=1}^R = \begin{cases} (y_i)_{i=1}^S, & i \in S \\ 0, & i \notin S \end{cases}$

De coup A peut être recouvert par $\sum_{r=0}^s \binom{d}{r} (2/s)^r \leq \frac{(2d/s)^{s+1}}{1 - 2d/s}$

Finallement $\mathcal{N}(s, A, \|\cdot\|_\infty) \leq (2d/s)^{s+1}$ compare to $(2/s)^d$

entropy

$$(s+1) \log d + (s+1) \log(2/s)$$

$$vs d \log(2/s)$$

sparsité cool!

Rôle de l'entropie en régression

Modèle de régression non paramétrique

$$\left\{ \begin{array}{l} X_1 \dots X_n \text{ iid copies de } X \in [0,1]^d \\ Y_i = f_0(X_i) + \varepsilon_i, \quad 1 \leq i \leq n \\ \varepsilon_i \text{ iid } \sim \mathcal{N}(0,1) \text{ indép des } X_i \\ f_0 : [0,1]^d \rightarrow [-M,M], \quad M > 0 \text{ inconnue} \end{array} \right.$$

Estimateur minimiseur de risque empirique

$$\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{arg\,min}} \left\{ \sum_{i=1}^n (Y_i - f(x_i))^2 \right\}$$

pour \mathcal{F} classe de fonctions : $\mathbb{R}^d \rightarrow [-M,M]$

Théorème : for all $S \in (0,1)$, $\varepsilon > 0$

$$\mathbb{E} \left[(\hat{f}(x) - f_0(x))^2 \right] \leq (1+\varepsilon) \left[\inf_{f \in \mathcal{F}} \mathbb{E} \left[(f(x) - f_0(x))^2 \right] + M^2 \frac{18 \log W + 72}{n \varepsilon} + 3\varepsilon S M \right]$$

biases variance

pour $W = W(S, \mathcal{F}, \| \cdot \|_\infty) \geq 3$ entropie sur un espace fonctionnel

2.2 Contrôle de l'entropie pour une classe de NN sparses

On considère $\mathcal{F}(L, N, s)$ la classe des NN ReLU avec

- * profondeur $L \geq 2$
- * vecteur $N = (N_l)_{0 \leq l \leq L}$

$N_0 = d$ dimension d'entrée
 N_l nombre de neurones
 pour $l > 0$ $1 \leq l \leq L-1$
 $N_L = 1$ dimension sortie

- * sparsité

$$\sum_{l=1}^L \|A_l\|_0 + \|b_l\|_0 \leq s$$

- * poids bornés

$$\max_{1 \leq l \leq L} (\|A_l\|_\infty, \|b_l\|_\infty) \leq 1$$



souvent utilisé comme initialisation
 avant optimisation \rightarrow correspond à l'usage

Théorème: pour $V = \sum_{\ell=0}^L (N_\ell + 1)$

on a $\forall \delta \in (0, 1)$,

$$\log \mathcal{N}(s, \mathcal{F}(L, N, \delta), \| \cdot \|_\infty) \leq (s+1) \log \left(\frac{2LV^2}{\delta} \right)$$

Preuve: idée on se ramène au cas de l'entropie de vecteurs sparses

Lemme outil:

Pour $f = R(\Phi)$, $f' = R(\Phi')$ $\in \mathcal{F}(L, N, s)$

avec poids de Φ et Φ' à distance au plus $\varepsilon > 0$

$$\text{Alors } \|f - f'\|_\infty \leq \varepsilon L V$$

mesure la propagation d'une erreur ε sur les poids
dans la valeur du résultat

Discussion

- * Fait aussi le borne sur la complexité \nearrow avec L, s, N
- * Le compromis biais variance sera fait précisément dans le chapitre 4

Notamment, choisir L très grand n'est pas forcément la meilleure solution

- * Il y a d'autres mesures de complexité comme la complexité de Rademacher

pour $A \subset \mathbb{R}^d$, $R(A) = \frac{1}{d} \mathbb{E}_\sigma \left[\sup_{a \in A} \sum_{i=1}^d \sigma_i a_i \right]$ avec $\sigma = (\sigma_i)_{1 \leq i \leq n}$ signes iid