

- Pairs d'(observation, classe) aléatoires:  
 $(X, Y) \in \mathbb{R}^d \times \{-1, 1\}$

- Classe de fonctions:  
 $\mathcal{F} = \{f : \mathbb{R}^d \rightarrow \mathbb{R}\}$

- Erreur:  $f(X)Y < 0$

- Erreur de généralisation:  
 $R(f) = P[f(X)Y < 0]$

- Échantillon iid:  $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$

- Algorithme d'apprentissage sort  $f_n \in \mathcal{F}$

- Objectif: avec une probabilité de au moins  $1 - \delta$   
 $R(f_n) \leq \text{Borne}(n, D_n, f_n, \mathcal{F})$

- Motivations

- purement théorique
- garantie pour la fiabilité statistique
- sélection de modèle
- Expliquer la capacité de généralisation des méthodes d'apprentissage existantes
- direction de développement des nouvelles méthodes

- Erreur empirique:  $\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n I_{\{f(X_i)Y_i < 0\}}$

- Le meilleur classifieur empirique:

$$f_n^* = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f)$$

- Avec un probabilité de au moins  $1 - \delta$

$$R(f_n^*) \leq \hat{R}_n(f_n^*) + O\left(\sqrt{\frac{\text{VC}_{\mathcal{F}} \log n}{n}}\right) + O\left(\sqrt{\frac{\log 1/\delta}{n}}\right)$$

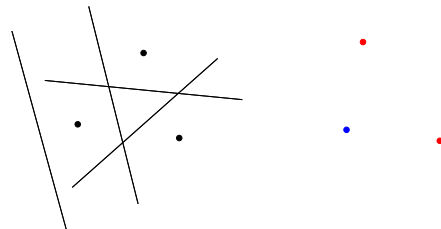
[Vapnik-Chervonenkis, '74]

- $\text{VC}_{\mathcal{F}}$  est la cardinalité de l'ensemble plus grand dans  $\mathbb{R}^d$  que  $\mathcal{F}$  peut "shatterer"

- $\mathcal{F}$  peut shatterer  $\{x_1, \dots, x_n\}$  si l'on peut apprendre n'importe quel classement  $\{y_1, \dots, y_n\}$  sans erreur

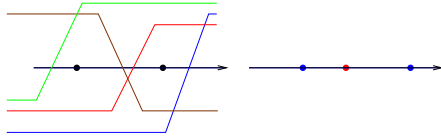
- Exemples

- Découpages linéaires en 2D:  $\text{VC}_{\mathcal{F}} = 3$



• Exemples

- Sigmoides en 1D:  $VC_{\mathcal{F}} = 2$



• Exemples

- $|\mathcal{F}|$  est fini:  $VC_{\mathcal{F}} \leq \log_2 |\mathcal{F}|$
- $\mathcal{F}$  est la classe des (hyper)rectangles dans  $\mathbb{R}^d$ :  $VC_{\mathcal{F}} = 2d$
- $\mathcal{F}$  est la classe des demi-espaces dans  $\mathbb{R}^d$  (classifieurs linéaires):  $VC_{\mathcal{F}} = d + 1$
- $\mathcal{F}$  est la classe des polygones convexes dans  $\mathbb{R}^2$ :  $VC_{\mathcal{F}} = \infty$
- $\mathcal{F}$  est la classe des combinaisons convexes des decision stumps dans  $\mathbb{R}^d$ :  $VC_{\mathcal{F}} = \infty$

•  $R(f_n^*) \leq \widehat{R}_n(f_n^*) + O\left(\sqrt{\frac{VC_{\mathcal{F}} \log n}{n}}\right)$

- $VC_{\mathcal{F}}$  est énorme (même  $\infty$ ) pour les classes de fonctions utilisées en pratique
- $VC_{\mathcal{F}}$  est indépendant des données: approche du pire des cas

• Bornes de type VC en générale:

$R(f_n) \leq ErrEmp(D_n, f_n) + Complexite(\mathcal{F}, n)$

• Objectif: complexités dépendantes des données

$R(f_n) \leq ErrEmp(D_n, f_n) + ComplEmp(\mathcal{F}, D_n, n)$

- $ComplEmp(\mathcal{F}, D_n, n) \ll Complexite(\mathcal{F}, n)$
- $ComplEmp(\mathcal{F}, D_n, n)$  peut être calculé

• Boosting [Freund, '95]

- Combinaisons convexes des fonctions de base simples:

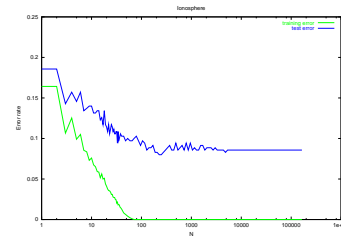
$$\mathcal{F} = \left\{ f(x) = \sum_{i=1}^N w_i h_i(x) : h_i \in \mathcal{H}, \sum_{i=1}^N w_i = 1 \right\}$$

• Exemple

- $\mathcal{H}$  est l'ensemble des decision stumps
- $VC_{\mathcal{F}} \xrightarrow{N \rightarrow \infty} \infty$

• Boosting [Freund, '95]

- En pratique  $R(f_n) \xrightarrow{N \rightarrow \infty} \infty$

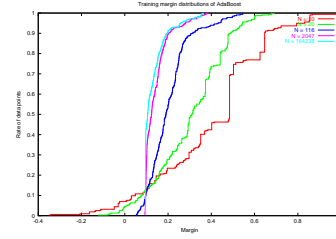


• Autres exemples

- Support vector machines [Cortes-Vapnik, '95]
- Réseaux de neurones avec des **petits poids** (early stopping, weight decay) [Bartlett, '98]

• Distribution de la **marge** d'entraînement

- Marge:  $f(x)y$



- Boosting **maximise la marge minimale**

- Erreur  $\gamma$  empirique:  $\widehat{R}_n^\gamma(f) = \frac{1}{n} \sum_{i=1}^n I_{\{f(x_i)y_i < \gamma\}}$

- Avec une probabilité de au moins  $1 - \delta$

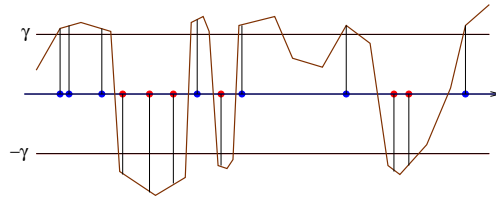
$$R(f_n^*) \leq \widehat{R}_n^\gamma(f_n^*) + O\left(\sqrt{\frac{\text{fat}_{\mathcal{F}}(\gamma) \log n}{n}}\right) + O\left(\sqrt{\frac{\log 1/\delta}{n}}\right)$$

[Schapire-Freund-Bartlett-Lee, '98]

- $\text{fat}_{\mathcal{F}}(\gamma)$  est la cardinalité de l'ensemble plus grand dans  $\mathbb{R}^d$  que  $\mathcal{F}$  peut " **$\gamma$ -shatterer**"

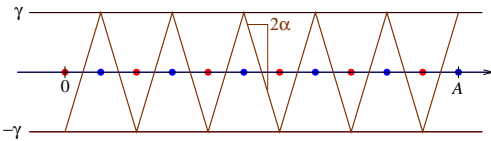
- $\mathcal{F}$  peut  $\gamma$ -shatterer  $\{x_1, \dots, x_n\}$  si l'on peut apprendre **n'importe quel** classement  $\{y_1, \dots, y_n\}$  avec une **marge**  $> \gamma$

- $\gamma$ -séparation:  $f(x_i)y_i \geq \gamma$



- Exemple:  $\mathcal{F} = \text{Lip}_{2\alpha}$  on  $[0, A]$

$$\text{fat}_{\mathcal{F}}(\gamma) = \left\lfloor \frac{A\alpha}{\gamma} \right\rfloor + 1$$



- $R(f_n^*) \leq \widehat{R}_n^\gamma(f_n^*) + O\left(\sqrt{\frac{\text{fat}_{\mathcal{F}}(\gamma) \log n}{n}}\right)$

- $\widehat{R}_n^\gamma(f) \geq \widehat{R}_n(f)$ ,  $\widehat{R}_n^\gamma(f) = \widehat{R}_n(f)$

- $\text{fat}_{\mathcal{F}}(\gamma) \leq \text{VC}_{\mathcal{F}}$ ,  $\text{fat}_{\mathcal{F}}(0) = \text{VC}_{\mathcal{F}}$

- la borne est valide **uniformément en  $\gamma$** : on peut choisir  $\gamma$  après que l'on ait vu les données

- $\text{fat}_{\mathcal{F}}$  est encore **indépendant des données**

- $\text{fat}_{\mathcal{F}}$  peut être encore **grand** (même  $\infty$ )

- Bornes de type fat en générale:

$$R(f_n) \leq \text{FatErrEmp}(D_n, f_n, \gamma) + \text{FatComplexite}(\mathcal{F}, n, \gamma)$$

- Objectif: complexités de **marge**, dépendantes des données

$$R(f_n) \leq \text{FatErrEmp}(D_n, f_n, \gamma) + \text{FatComplEmp}(\mathcal{F}, D_n, n, \gamma)$$

- $\text{FatComplEmp}(\mathcal{F}, D_n, n, \gamma) \ll \text{FatComplexite}(\mathcal{F}, n, \gamma)$
- $\text{FatComplEmp}(\mathcal{F}, D_n, n, \gamma)$  peut être calculé

- Avec une probabilité de au moins  $1 - \delta$

$$R(f_n^*) \leq \widehat{R}_n^*(f_n^*) + O\left(\sqrt{\frac{\text{fat}_{\mathcal{F}, D_n}(\gamma) \log n}{n}}\right) + O\left(\sqrt{\frac{\log 1/\delta}{n}}\right)$$

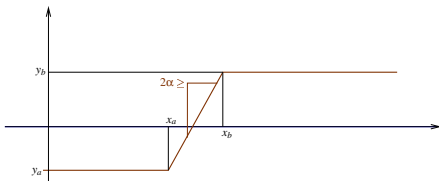
- $\text{fat}_{\mathcal{F}, D_n}(\gamma)$  est la cardinalité du sous-ensemble plus grand de  $D_n$  que  $\mathcal{F}$  peut  $\gamma$ -shatterer
- trouver  $\text{fat}_{\mathcal{F}, D_n}(\gamma)$  n'est plus un problème combinatoire mais **algorithmique**
- $\text{fat}_{\mathcal{F}, D_n}(\gamma)$  dépend des points  $x_i$  mais ne dépend pas des étiquettes  $y_i$

Mesurer la dimension fat-shattering empirique 21

- Exemple:

- sigmoïdes 1D, **linéaire par morceaux**:

$$g^{(x_a, x_b, y_a, y_b)}(x) = \begin{cases} y_a & \text{if } x \leq x_a \\ y_b & \text{if } x \geq x_b \\ y_a + \frac{y_b - y_a}{x_b - x_a}(x - x_a) & \text{sinon} \end{cases}$$



Mesurer la dimension fat-shattering empirique 22

- Exemple:

- famille des sigmoïdes avec une **penne bornée**:

$$G_\alpha = \left\{ g^{(x_a, x_b, y_a, y_b)} : \left| \frac{y_b - y_a}{x_b - x_a} \right| \leq 2\alpha \right\}$$

- famille des **combinaisons convexes**:

$$F_\alpha = \left\{ f(x) = \sum_{i=1}^N w_i g_i(x) : g_i \in G_\alpha, \sum_{i=1}^N w_i = 1 \right\}$$

Mesurer la dimension fat-shattering empirique 23

- Problème **algorithmique**

- trouver (la cardinalité) du **plus grand sous-ensemble** de  $X_n = \{x_1, \dots, x_n\}$  ( $x_i < x_{i+1}$ ) que  $F_\alpha$  peut  $\gamma$ -shatterer

- Lemme

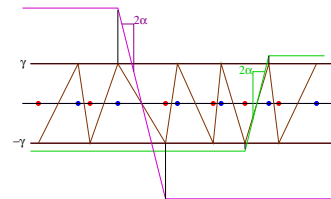
- $X_n$  est  $\gamma$ -shatteré par  $F_\alpha$  ssi

$$\sum_{i=1}^n \frac{1}{d_i} \leq \frac{\alpha}{\gamma}$$

où  $d_i = x_i - x_{i-1}$ .

Mesurer la dimension fat-shattering empirique 24

- Preuve



- $w_i = \frac{1}{d_i} / \sum_{j=1}^n \frac{1}{d_j}$

• Problème dual:

- trouver une sous-séquence  $(x_{j_1}, \dots, x_{j_k})$  qui minimise le coût

$$\sum_{i=1}^{k-1} \frac{1}{d_{j_i, j_{i+1}}}$$

parmi toutes les sous-séquences de longueur  $k$ .

Mesurer la dimension fat-shattering empirique 27

- Trouver la dimension fat-shattering empirique  $\text{fat}_{\mathcal{F}_{\alpha, X_n}}^{\gamma}(\gamma)$

- arrêter quand  $C(k; 1, n) > \frac{\alpha}{\gamma} \rightarrow$  retourner  $k$
- temps de  $O(n^2 \text{fat}_{\mathcal{F}_{\alpha, X_n}})$  ( $O(n^3)$  maximum)

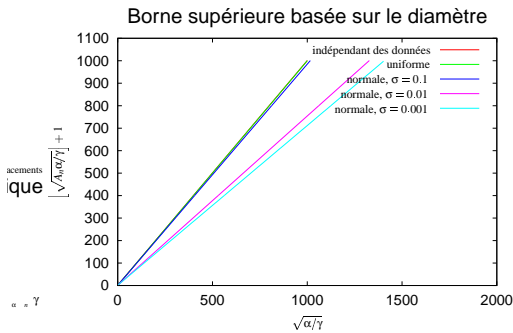
• Fat-shattering dans le pire des cas:

- $\text{fat}_{\mathcal{F}_{\alpha}}(\gamma) = \infty$  si la distribution de données n'est pas bornée
- $\text{fat}_{\mathcal{F}_{\alpha}}(\gamma) = \lfloor \sqrt{A\alpha/\gamma} \rfloor + 1$  si les données sont dans  $[a, a + A]$

• Fat-shattering en utilisant le diamètre empirique:

- $A_n = \max_i x_i - \min_i x_i$
- $\text{fat}_{\mathcal{F}_{\alpha, X_n}}^{\gamma}(\gamma) \leq \lfloor \sqrt{A_n \alpha / \gamma} \rfloor + 1$

Mesurer la dimension fat-shattering empirique 29



• Solution:

- sous-séquence optimale de longueur  $k + 1$  entre  $x_p$  et  $x_r$ :

$$S(k; p, r) = (x_p = x_{j_1}, \dots, x_{j_{k+1}} = x_r)$$

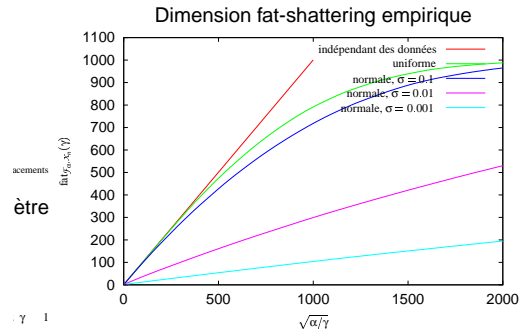
- coût de  $S(k; p, r)$ :  $C(k; p, r) = \sum_{i=1}^k \frac{1}{d_{j_i, j_{i+1}}}$

• définition récursive:

$$C(k; p, r) = \begin{cases} \frac{1}{d_{p,r}} & \text{if } k = 1 \\ \min_{q: p+k-1 \leq q \leq r-1} (C(k-1; p, q) + C(1; q, r)) & \text{if } k > 1. \end{cases}$$

- programmation dynamique: calculer  $C(1; 1, n), C(2; 1, n), \dots, C(k; 1, n)$  en temps  $O(n^2 k)$

Mesurer la dimension fat-shattering empirique 28



Mesurer la dimension fat-shattering empirique 30

• Exemple:

- $\mathcal{F}_{\alpha} = \text{Lip}_{2\alpha}$   $d$ -dimensionnel
- $\mathcal{F}_{\alpha}$  ne peut pas  $\gamma$ -shatterer  $X_n$  si deux points sont plus proches que  $\frac{\gamma}{\alpha}$
- la dimension fat-shattering dans le pire des cas:  $\frac{\gamma}{\alpha}$ -packing number
- définir le graphe  $G_{\alpha, \gamma}(V, E)$

$$V = X_n$$

$$E = \left\{ (x_i, x_j) : \|x_i - x_j\| \leq \frac{\gamma}{\alpha} \right\}$$

- trouver  $\text{fat}_{\mathcal{F}_{\alpha, X_n}}^{\gamma}(\gamma) \equiv$  trouver un ensemble maximal des sommets indépendants dans  $G_{\alpha, \gamma}$

• Idée [Vapnik et al. 1994]:

- mesurer l'erreur empirique **séparément sur les deux moitiés** des données

$$\widehat{R}_{n/2}^{(1)}(f) = \frac{2}{n} \sum_{i=1}^{n/2} I_{\{f(X_i)Y_i < 0\}}$$

$$\widehat{R}_{n/2}^{(2)}(f) = \frac{2}{n} \sum_{i=n/2+1}^n I_{\{f(X_i)Y_i < 0\}}$$

- la complexité **effective** (dépendante des données) de  $\mathcal{F}$  est mesurée par la **discrédence maximale**

$$\max_{f \in \mathcal{F}} \left( \widehat{R}_{n/2}^{(2)}(f) - \widehat{R}_{n/2}^{(1)}(f) \right)$$

• Combiner avec des bornes basées sur la **marge**:

- $\widehat{R}_{n/2}^{\gamma}(f) = \frac{2}{n} \sum_{i=1}^{n/2} I_{\{f(X_i)Y_i < \gamma\}}$

- avec une probabilité de au moins  $1 - \delta$

$$R(f_n^*) \leq \widehat{R}_n^{\gamma}(f_n^*) + \max_{f \in \mathcal{F}} \left( \widehat{R}_{n/2}^{(2)}(f) - \widehat{R}_{n/2}^{\gamma}(f) \right) + O\left(\sqrt{\frac{\log 1/\delta}{n}}\right)$$

Autres complexités empiriques 35

• Complexité de **VC empirique**

• Complexité de **Rademacher/gaussienne** [Bartlett, '01], [Koltchinskii–Panchenko–Lozano]:

- apprendre la donnée avec des **étiquettes aléatoires**

• [Koltchinskii–Panchenko, '00]:

- boosting n'utilise pas **tout l'espace** de combinaison convexe
- seulement un sous-espace de complexité **réduite**
- engendré par les fonctions de base avec des **plus grands poids**

• Idée [Vapnik et al. 1994]:

- soit  $D_n^* = \{(X_1, Y_1), \dots, (X_{n/2}, Y_{n/2}), (X_{n/2+1}, -Y_{n/2+1}), \dots, (X_n, -Y_n)\}$

- $\widehat{R}_{n/2}^{(2)}(f) = 1 - \widehat{R}_{n/2}^{(1)}(f)$

- $\arg \max_{f \in \mathcal{F}} \left( \widehat{R}_{n/2}^{(2)}(f) - \widehat{R}_{n/2}^{(1)}(f) \right) = \arg \min_{f \in \mathcal{F}} \widehat{R}_n(f)$

• Combiner avec des bornes basées sur la **marge**:

- **renverser** les étiquettes

$$\max_{f \in \mathcal{F}} \left( \widehat{R}_{n/2}^{(2)}(f) - \widehat{R}_{n/2}^{\gamma}(f) \right) = \max_{f \in \mathcal{F}} \left( 1 - \widehat{R}_{n/2}^{(2)}(f) - \widehat{R}_{n/2}^{\gamma}(f) \right)$$

$$= 1 - \min_{f \in \mathcal{F}} \left( \widehat{R}_{n/2}^{(2)}(f) + \widehat{R}_{n/2}^{\gamma}(f) \right)$$

- minimiser  $\left( \widehat{R}_{n/2}^{(2)}(f) + \widehat{R}_{n/2}^{\gamma}(f) \right)$  est un **problème algorithmique** (pas très difficile)
- minimiser  $\left( \widehat{R}_{n/2}^{(2)}(f) + \widehat{R}_{n/2}^{\gamma}(f) \right)$  au meilleur  $\gamma$  n'est pas facile