Apprentissage Automatique: Théorie de l'apprentissage

S. Herbin, A. Chan Hon Tong

stephane.herbin@onera.fr

Introduction

Cours précédents

- Principes généraux d'apprentissage : données apprentissage/validation/test, optimisation, évaluation, régularisation
- Plusieurs algorithmes de *classification supervisée* : plus proche voisin, classifieur Bayésien, machine à vecteurs de support (SVM), Réseaux de neurones.

Objectifs de ce cours

- Pourquoi ça marche : justifications théoriques, comment ça s'exprime.
- Domaine : « Statistical Machine Learning » ou « Computational Learning Theory »

Eléments de théorie de l'apprentissage statistique

Minimisation du risque empirique

PAC learning

Caractériser les familles de prédicteurs

Minimisation du risque empirique

Rappels

f une fonction de prédiction $\mathcal{X} \to \mathcal{Y}$
Un ensemble de données $D_n = \{(x_i, y_i)\}_{1 \leq i \leq n}$ issues d'une distribution P

Plusieurs fonctions d'erreur

- Coût : $l(y, y') \in [0, +\infty[$, par exemple $l_{01}(y, y') = \mathbb{1}_{\{y \neq y'\}}$
- Risque vrai ou réel :

$$L(f) = E[l(f(X), Y)] = \int l(f(x), y) dP(x, y)$$

- Risque empirique :

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), Y_i)$$

Minimisation du risque empirique (MRE)

Principe

- On se donne une famille de prédicteurs $\mathcal{F}$ (on parle parfois d'hypothèses)
- Algorithme = Trouver dans cette famille celui qui minimise le risque empirique :

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), Y_i)$$

Erreur de généralisation

- Le risque de ce prédicteur est potentiellement supérieur au risque réel (erreur de Bayes)

$$L(\hat{f}_n) \gg L^*$$

où $L^* = \inf_f L$ est le risque réel (idéal)

Structure des erreurs

Autre décomposition biais/variance

$$L(\hat{f}_n) - L^* = \underbrace{L(\hat{f}_n) - L(f^*)}_{\text{estimation, stochastique}} + \underbrace{L(f^*) - L^*}_{\text{approximation, déterministe}}$$

où $f^* = \arg\min_{f \in \mathcal{F}} L(f)$ est le meilleur prédicteur possible de la famille $\mathcal{F}$.
Deux sources d'erreur :

- Le nombre limité de données
- La famille des prédicteurs (arbres, réseaux de neurones...)
- L'erreur d'approximation est liée à la modélisation du problème
- L'erreur d'estimation est liée à l'apprentissage : c'est elle que l'on peut essayer de contrôler.

Que veut-dire apprendre ?

Questions

1. Comment garantir que mon algorithme d'apprentissage se comporte bien : $L(\hat{f}_n) \xrightarrow[n\to\infty]{} L(f^*)$?
2. Combien de données pour garantir une erreur de généralisation minimale ?
3. Comment contrôler ou caractériser l'espace des fonctions de prédiction $\mathcal{F}$?

PAC learning

Comment qualifier un algorithme d'apprentissage ?

Un premier cas simplifié

- Classification binaire : $\mathcal{Y} = \{0, 1\}$
- $|\mathcal{F}|$ fini
- $\mathcal{F}$ contient un prédicteur parfait ($L(f^*) = 0$) : on parle de problème « réalisable »

Quelques conséquences

- Pour tout échantillon D_n : $L_n(f^*) = 0$ et $L_n(\hat{f}_n) = 0$ car $l(y, y') \geq 0$
- Mais en général $L(\hat{f}_n) > 0$ (erreur de généralisation)

Objectif

- Majorer la probabilité de faire des erreurs d'au plus ϵ : $P[L(\hat{f}_n) > \epsilon]$ par une fonction de ϵ et n.

Démonstration

On va prouver que

$$P[L(\hat{f}_n) > \epsilon] \leq |\mathcal{F}|e^{-n\epsilon}$$

Étapes

1. On s'intéresse aux ensembles de données D_n *trompeurs*, c-à-d qui estiment un prédicteur $\hat{f}_n$ tel que $L(\hat{f}_n) > \epsilon$ (erreur réelle) mais avec $L_n(\hat{f}_n) = 0$ (pas d'erreur empirique)
2. Ces ensembles sont les seules sources d'erreur !
3. On va repérer ces ensembles à partir des prédicteurs erronés (ceux pour lesquels $L(f) > \epsilon$)
4. Puis on va calculer pour chaque prédicteur erroné la probabilité de tomber sur un ensemble trompeur

Détails I

1. On veut majorer la probabilité qu'un ensemble de données soit source d'erreur : $P\{D_n \ : \ L(\hat{f}_n) > \epsilon\}$
2. On repère les prédicteurs erronés : $\mathcal{F}_\epsilon = \{f \ : \ L(f) > \epsilon\}$
3. Les données source d'erreur sont trompeuses :

$$\{D_n : L(\hat{f}_n) > \epsilon\} \subset \bigcup_{f \in \mathcal{F}_\epsilon} \{D_n : L_n(f) = 0\}$$

4. On passe aux probabilités et on applique l'inégalité de Boole (« Union Bound ») :

$$P[L(\hat{f}_n) > \epsilon] \leq P[\bigvee_{f \in \mathcal{F}_\epsilon} \{L_n(f) = 0\}] \leq \sum_{f \in \mathcal{F}_\epsilon} P[L_n(f) = 0]$$

Détails II

5. On calcule la probabilité d'être erroné pour un ensemble $D_n = \{(x_i, y_i)\}_{1 \leq i \leq n}$:

$$\begin{aligned} P[L_n(f) = 0] &= P[\forall i \, f(x_i) = f^*(x_i)] \\ &= \prod_{1 \leq i \leq n} P[f(x) = f^*(x)] \\ &\leq \prod_{1 \leq i \leq n} (1 - \epsilon) = (1 - \epsilon)^n \leq e^{-n\epsilon} \end{aligned}$$

   car $f \in \mathcal{F}_\epsilon$

6. Et finalement

$$\begin{aligned} P[L(\hat{f}_n) > \epsilon] &\leq \sum_{f \in \mathcal{F}_\epsilon} e^{-n\epsilon} \\ &\leq |\mathcal{F}_\epsilon| e^{-n\epsilon} \leq |\mathcal{F}| e^{-n\epsilon} \end{aligned}$$

PAC learning

Interprétation de $P[L(\hat{f}_n) > \epsilon] \leq |\mathcal{F}|e^{-n\epsilon}$

- Cette inégalité est valable pour l'utilisation de l'algorithme MRE et pour toute famille finie de prédicteurs $\mathcal{F}$ qui contient un prédicteur sans erreur ($L(f^*) = 0$).
- Elle indique que l'on peut obtenir un prédicteur par MRE avec une probabilité d'erreur bornée.
- On peut reformuler le résultat : Si $n \geq \frac{\log(|\mathcal{F}|/\delta)}{\epsilon} = m(\epsilon, \delta)$ alors on aura un risque réel inférieur à ϵ avec une probabilité $1 - \delta$.
- La valeur de $m(\epsilon, \delta)$ ne dépend pas de P (la distribution de données) !

« Probably Approximately Correct learnable »

Définition

Une famille de prédicteurs $\mathcal{F}$ est PAC apprenable s'il existe une fonction $m :]0,1[^2 \rightarrow \mathbb{N}$ et un algorithme d'apprentissage tels que : pour tout $\epsilon > 0$ et $\delta > 0$, en appliquant l'algorithme d'apprentissage sur un échantillon de taille $m(\epsilon, \delta)$, on obtienne un prédicteur de risque inférieur à ϵ avec une probabilité de $1 - \delta$.

$$n \geq m(\epsilon, \delta) \Rightarrow P[L(\hat{f}_n) \leq \epsilon] \geq 1 - \delta$$

Relâcher la contrainte de réalisabilité

- Il est difficile de garantir : $\min_{f \in \mathcal{F}} L(f) = 0$ et donc de garantir une borne absolue sur le risque
- Lorsque $\min_{f \in \mathcal{F}} L(f) > 0$, on cherche plutôt à borner l'erreur d'estimation du risque réel : $L(\hat{f}_n) - L(f^*)$
- Cela conduit à une version dérivée de la notion de « PAC apprenable » dite « agnostique » = on ne sait pas si le prédicteur idéal est dans $\mathcal{F}$.
- On peut montrer, avec probabilité $1 - \delta$:

$$L(\hat{f}_n) - L(f^*) \leq \sqrt{\frac{2\log(2|\mathcal{F}|/\delta)}{n}}$$

- Ce qui donne comme indicateur de complexité PAC :

$$m_{\text{agnostique}}(\epsilon, \delta) = \frac{2\log(2|\mathcal{F}|/\delta)}{\epsilon^2}$$

Eléments de démonstration

Inégalité de Hoeffding

- Si $Z_1, Z_2 \dots Z_n$ sont des variables i.i.d. à valeur dans $[0, 1]$.
- Alors pour tout $\epsilon > 0$:

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - E(Z_1)\right| > \epsilon\right] \leq 2\exp(-2n\epsilon^2)$$

Convergence uniforme

- On s'intéresse au comportement de tous les prédicteurs et on utilise l'inégalité :

$$L(\hat{f}_n) - L(f^*) \leq 2\sup_{f\in\mathcal{F}} |L(f) - L_n(f)|$$

PAC learning

Des résultats généraux

- Permet de garantir que l'apprentissage MRE fonctionne pour des familles finies de prédicteurs
- Donne des bornes sur le nombre de données utiles (indépendamment de la distribution sous-jacente !)

Mais des limitations

- Ne donne de résultats que pour les familles finies de prédicteurs
- Ne dit rien sur la nature des prédicteurs
- Ne dit rien sur l'erreur d'approximation $L(f^*) - L^*$
- Calculer le MRE peut être complexe si $|\mathcal{F}|$ est grand
- Les bornes sont grossières (ne dépendent ni des données, ni de la nature des prédicteurs)

Caractériser les familles de prédicteurs

Des résultats plus fins ?

Que faire lorsque $|\mathcal{F}| = \infty$?

- Beaucoup de familles de prédicteurs sont dans ce cas (arbres, réseaux de neurones, bayésien naïf, ...)
- Ils ont une expressivité variable et contrôlable (profondeur des arbres, couches des RN)
- Peut-on produire des résultats comparables au cas fini ?

Comment particulariser le « PAC learning » ?

- Les résultats précédents ne dépendent pas de la nature de $\mathcal{F}$
- Comparer différentes familles de prédicteurs ?

➪ Théorie de Vapnik Chervonenkis [Vapnik, 2013]

Complexité des familles de prédicteurs

Intuitions

- Expressivité = Capacité à discriminer un grand *nombre* de données.
- Famille de prédicteurs décrites par un nombre fini de paramètres : complexité = # paramètres ?. . . pas tout à fait.

Un problème combinatoire

- Idée = on compte le nombre maximal de prédicteurs capables de discriminer un jeu de données quelconque de taille m.
- On regarde la forme de la fonction de croissance :

$$G_{\mathcal{F}}(n) = \max_{x_1,x_2,\ldots,x_n \in \mathcal{X}} |\{(f(x_1), f(x_2), \ldots f(x_n)) : f \in \mathcal{F}\}|$$

- On a nécessairement : $G_{\mathcal{F}}(n) \leq 2^n$

Fonction de croissance

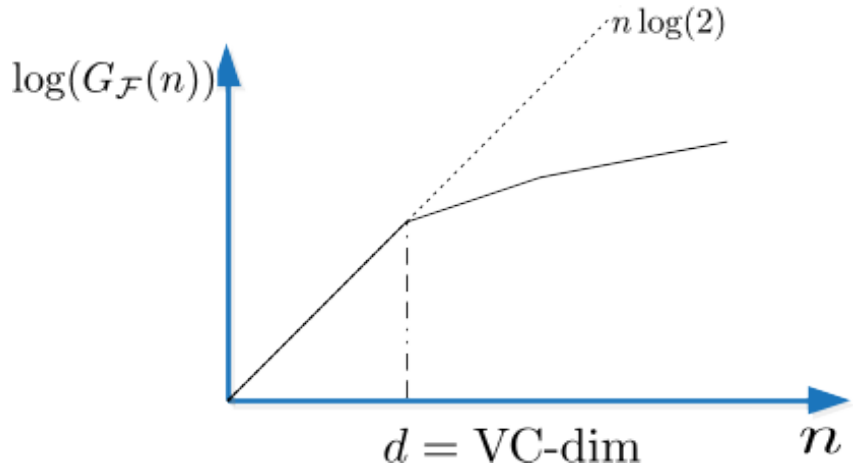

Figure 1 – Allure de la fonction de croissance.

Allure

- La fonction croit comme 2^n jusqu'à une certaine valeur (h sur Figure 1) puis moins vite
- Cette valeur est la dimension de Vapnik-Chervonenkis
- ! ! ! Elle peut être infinie ! = famille de prédicteurs très complexe ➪ sur-apprentissage (en fait, aucune garantie d'apprentissage)

Dimension de Vapnik-Chervonenkis

Définition

- Plus grande taille n de données $X_n = (x_1, x_2, \ldots, x_n)$ étiquetée de manière quelconque qui puisse être discriminée par un élément de $\mathcal{F}$
- On dit que $\mathcal{F}$ pulvérise X_n
- Conséquence : il suffit de trouver une configuration de n points pulvérisée pour avoir VC-dim$(\mathcal{F}) \geq n$

Propriétés

- Lien avec fonction de croissance :
  VC-dim$(\mathcal{F}) = \max_n \{n : G_{\mathcal{F}}(n) = 2^n\}$
- Lemme de Sauer : si VC-dim$(\mathcal{F}) = d < \infty$ alors
  $G_{\mathcal{F}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$.
  En particulier, si $n > d+1$, $G_{\mathcal{F}}(n) \leq (e.n/d)^d$ (croissance polynomiale < exponentielle)

Un exemple 2D : hyperplans dans $\mathbb{R}^2$

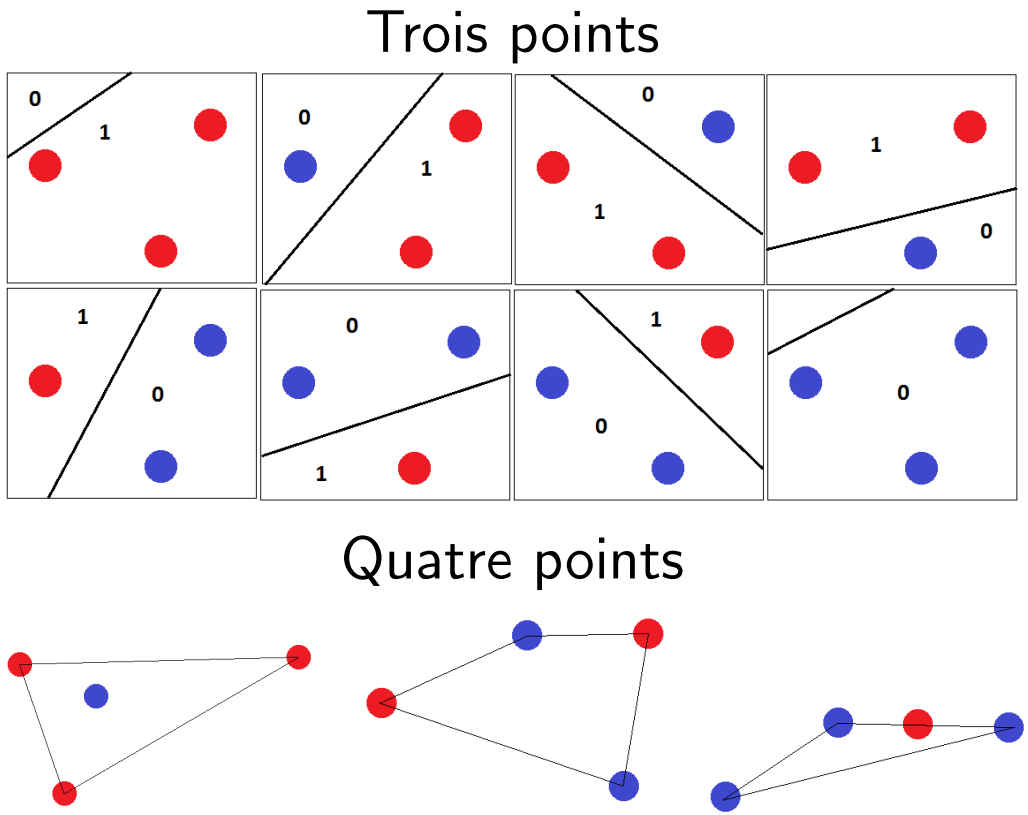

Figure 2 – Pulvérisation par hyperplan.

Les configurations de 4 points ne sont pas linéairement séparables :
➪ VC-dim(« Hyperplans ») $= 3$

Un exemple 2D : rectangles dans $\mathbb{R}^2$

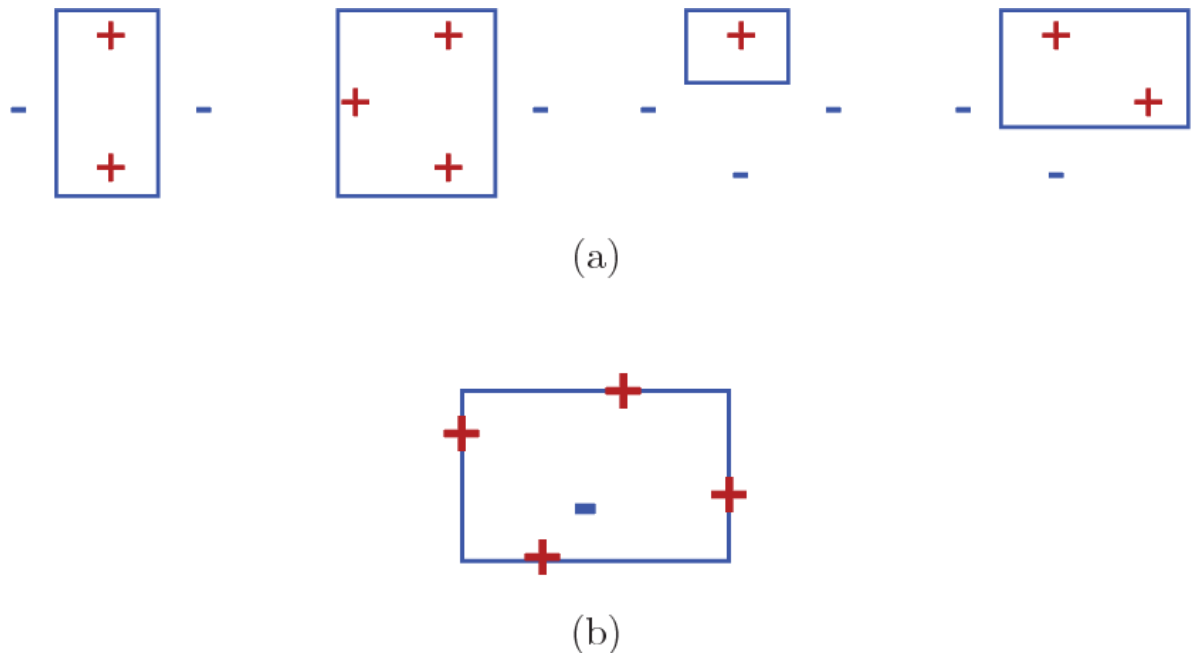

Figure 3 – Pulvérisation par fonction rectangle.

Pas moyen de séparer l'étiquetage (b) avec des rectangles :
➪ VC-dim(« Rectangles ») $= 4$

Un autre exemple 1D : fonction caractéristique sinusoïdale

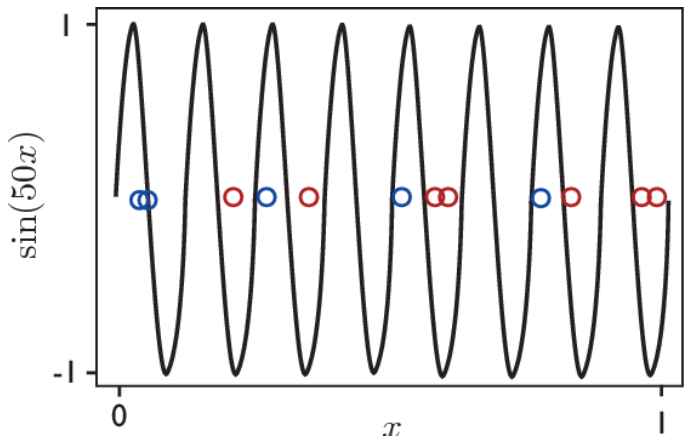

Figure 4 – Pulvérisation par fonction sinusoïdale.

- $f(x) = \text{signe}(\sin(\omega x))$ où $\omega \in [0, 2\pi)$.
- On peut trouver un ω qui sépare un ensemble de points de taille n quelconque
  - $x_j = 2\pi 10^{-j}$
  - $w = \frac{1}{2}\left(1 + \sum_{i=1}^{n} \frac{1-y_i}{2} 10^i\right)$

➪ VC-dim(« Sinusoides ») $= \infty$

Complexité et erreur d'estimation

On peut montrer, si VC-dim$(\mathcal{F}) = d$, avec probabilité $1 - \delta$:

$$L(\hat{f}_n) \leq \inf_{f \in \mathcal{F}} L(f) + \sqrt{\frac{2d(1 + \log(n/d))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Interprétation

- Si la famille de prédicteurs est de dimension VC fini, alors elle est PAC apprenable
- On peut borner l'erreur sur le risque par un $O\left(\sqrt{\frac{\log(n/d)}{n/d}}\right)$
- On peut aussi montrer que si la dimension VC de $\mathcal{F}$ est infinie, elle n'est pas PAC apprenable.

Ce *théorème fondamental de l'apprentissage statistique* implique qu'il y a **équivalence entre PAC apprenable et avoir une dimension VC finie** pour une famille de prédicteurs.

Exemples de dimensions VC

- Hyperplans dans $\mathbb{R}^d$: $d+1$
- Rectangles alignés sur les axes dans $\mathbb{R}^2$: 4
- Rectangles quelconques dans $\mathbb{R}^2$: 7
- Triangles dans $\mathbb{R}^2$: 7
- Polygones convexes dans $\mathbb{R}^2$: ∞
- Réseaux de neurones avec RELU (W paramètres et L couches)[Bartlett et al., 2019] : $\geq c \cdot WL\log(W/L)$ et $\leq C \cdot WL\log W$

D'autres inégalités I

Fonction de croissance

- On peut également montrer, avec probabilité $1-\delta$, $\forall f \in \mathcal{F}$:

$$L(f) \leq L_n(f) + \sqrt{\frac{2G_{\mathcal{F}}(n)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- ou de manière plus générale

$$P\left[|L(f) - L_n(f)| > \epsilon\right] \leq 4 \cdot G_{\mathcal{F}}(2n) \exp\left(-n\epsilon^2/8\right)$$

- La difficulté est d'estimer la fonction de croissance (la dimension VC est une simplification)

D'autres inégalités II

Complexité de Rademacher

- Définition : espérance de la « pire mauvaise classification »

$$R_n(\mathcal{F}, D_n) = E_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right] \text{ et } \bar{R}_n(\mathcal{F}) = E_P[R_n(\mathcal{F}, D_n)]$$

  où σ_i est une variable aléatoire uniforme i.i.d. sur $\{-1, 1\}$

- C'est une quantité qui dépend de la distribution
  ➪ bornes plus fines

D'autres inégalités III

Complexité de Rademacher

- Lien avec fonction de croissance

$$\bar{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log(G_{\mathcal{F}}(n))}{n}}$$

- Lien avec erreur d'estimation

$$L(f) \leq L_n(f) + \bar{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Les deux combinés redonnent la borne utilisant la VC-dimension.

MRE : bornes sur le nombre de données

Hypothèses

- $\mathcal{F}$ est fini ou sa dimension VC-dim$(\mathcal{F}) = d$ est finie.
- Fonction de coût $0-1$ $(l(y, y') = \mathbb{1}_{\{y \neq y'\}})$.
- Inégalité à probabilité $1 - \delta$.

$\lvert\mathcal{F}\rvert$	Réalisable $(\inf_{f\in\mathcal{F}} = 0)$ $P[L(\hat{f}_n) \leq \epsilon]$	Agnostique $(\inf_{f\in\mathcal{F}} > 0)$ $P[\lvert L(\hat{f}_n) - \inf_{f\in\mathcal{F}} \rvert \leq \epsilon]$
$< \infty$	$n \geq \dfrac{\log(\lvert\mathcal{F}\rvert/\delta)}{\epsilon}$	$n \geq \dfrac{2\log(2\lvert\mathcal{F}\rvert/\delta)}{\epsilon^2}$
$= \infty$	$n = O\left(\dfrac{d\log(1/\epsilon) + \log(1/\delta)}{\epsilon}\right)$	$n = O\left(\dfrac{d + \log(1/\delta)}{\epsilon^2}\right)$

Il y a encore bien d'autres questions...

- Autres algorithmes que MRE : arbres, ensembles, SVM, RN ?
- Comment introduire l'optimisation ($\arg\min$) dans les bornes
- Bornes inférieures (conditions nécessaires)
- Bornes dépendant des algorithmes, des distributions
- Contrôler la variance des écarts plutôt que $\sup$
- Utiliser les bornes en pratique pour contrôler les algorithmes
- Quel algorithme/stratégie utiliser avec des familles de dimension VC infinies ?
- Les algorithmes itératifs/séquentiels (renforcement, bandit...) : quelles garanties de convergence ?
- Pourquoi (et comment) les réseaux profonds qui contiennent plus de paramètres que de données généralisent-ils ?
- Pourquoi existe-t-il des exemples adversariaux ? Comment les contrer ?
- ...

Minimisation du risque empirique : Résumé

Résultats

+ Bornes théoriques
+ Justification de la faisabilité de l'apprentissage
+ Bornes indépendantes des distributions
- Résultats en probabilité $(1 - \delta)$
- Il y a d'autres algorithmes que MRE

Utilisations

+ Garantie
- Bornes trop lâches ou trop générales (convergence uniforme)
- Complexités difficiles à calculer (VC ou Rademacher)

Références I

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019).
Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks.
Journal of Machine Learning Research, 20(63) :1–17.

Devroye, L., Györfi, L., and Lugosi, G. (2013).
A probabilistic theory of pattern recognition, volume 31.
Springer Science & Business Media.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018).
Foundations of machine learning.
MIT press.

Shalev-Shwartz, S. and Ben-David, S. (2014).
Understanding machine learning : From theory to algorithms.
Cambridge university press.

Vapnik, V. (2013).
The nature of statistical learning theory.
Springer science & business media.