Les bases théoriques

Laurent Miclet

IRISA (Lannion)
France

EMINES 2024

Plan

L'île des mangayes

Vous arrivez sur une île isolée, dont la seule nourriture est la mangaye, un fruit dont vous ne connaissez rien.
Il vous faut rapporter du marché des mangayes dont le goût vous plaît.
Pour commencer, il vous faut trouver des indices pour guider votre choix.

Vous faites des expériences, jour après jour, et vous concluez que deux mesures sont significatives.
La couleur : une mangaye à la peau verte est amère, une mangaye à la peau marron est pourrie.
La consistance : trop dure , elle n'est pas assez mûre, acide ; trop molle, elle est blète, ecœurante.

NB : Ces deux caractéristiques sont plus ou moins corrélées.

Le problème

A partir de vos expériences, comment allez-vous choisir de bonnes mangayes pour vous nourrir ?

Fabrication d'un ensemble d'apprentissage

Sur quoi porte l'apprentissage ?

Sur le goût des mangayes et fonction de leur couleur et de leur consistance.

NB : Toute considération nutritive est écartée : on ne traite ici que de votre appétit.

Quelles sont vos connaissances ?

Le domaine d'étude est l'ensemble $\mathcal{X}$ des mangayes du marché.
Chaque mangaye est caractérisée par sa couleur et sa consistance.
Un vecteur x réel de dimension 2, dans l'espace de représentation $\mathcal{X} = \mathbb{R}^2$.

Construction de l'ensemble d'apprentissage

Un exemple d'apprentissage (*instance*) est une certaine mangaye, achetée au marché et dégustée ensuite.
Son goût (bon ou mauvais) lui donne une étiquette, un label, une classe.
On note $\mathcal{Y} = \{\omega_0, \omega_1\}$ ou simplement $\{0, 1\}$ l'ensemble de deux classes.
Un *ensemble d'apprentissage* (training set) S composé d'exemples (x, y).

NB : Au sens strict, S est un multi-ensemble.

Plan

Les hypothèses statistiques

L'environnement

Les mangayes rencontrées sont « engendrées » par une certaine distribution de probabilités $\mathcal{D}$ qui représente l'environnement local.

Vous ignorez cette distribution $\mathcal{D}$ qui commande les deux mesures.

Les étiquettes

Les étiquettes sont « engendrées » par une certaine fonction $f : \mathcal{X} \to \mathcal{Y}$

C'est cette fonction ou classificateur cible f que vous cherchez à apprendre par l'expérience.
Par un procédé d'apprentissage A (pour le moment inconnu), vous produisez un classificateur h qui est fonction de S.

Mesurer le succès de votre apprentissage

Erreur

L'erreur commise par un classificateur h est la probabilité que h ne produise pas la bonne étiquette sur un élément de $\mathcal{X}$ engendré par $\mathcal{D}$.

NB : Hypothèse de stationnarité sur $\mathcal{D}$

Expression de l'erreur

- Soit $\mathcal{A} \subset \mathcal{X}$. Alors, $\mathcal{D}$ fournit une valeur $\mathcal{D}(\mathcal{A})$ indiquant la probabilité d'observer un point (une mangaye) $x \in \mathcal{A}$.
- On peut exprimer $\mathcal{A}$ par la fonction $\pi : \mathcal{X} \to \{0, 1\}$, avec $\mathcal{A} = \{x \in \mathcal{X} : \pi(x) = 1\}$ et l'expression $\mathbb{P}_{x \sim \mathcal{D}}[\pi(x)]$ dénotera $\mathcal{D}(\mathcal{A})$.
- L'erreur $L_{\mathcal{D},f}(h)$ de la règle de prédiction h est donc :

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}(\{x \in \mathcal{X} : h(x) \neq f(x)\})$$

Plan

Minimiser l'erreur sur S

- S est produit selon $\mathcal{D}$ et étiqueté par la fonction cible f.
- On donne S en entrée à un algorithme A (un apprenant)
- A produit une hypothèse de classification h_S appartenant à un ensemble de fonctions possibles $\mathcal{H}$ dans lequel il travaille.
- A peut simplement explorer $\mathcal{H}$ pour y trouver h_S qui minimise l'erreur $L_S(h_S)$ sur S.

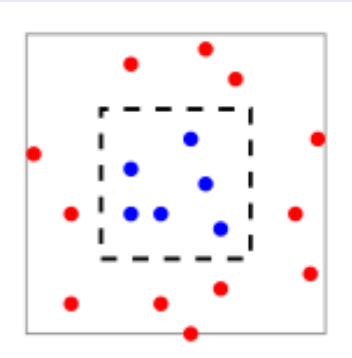
$S = ((x_1, y_1) \cdots (x_m, y_m))$

$$L_S(h_S) = \frac{|i \in \{1, m\} : h(x_i) \neq y_i|}{m}$$

Erreur empirique, risque empirique, perte empirique.
Minimisation de l'erreur empirique (ERM)

Surapprentissage

Exemple de distribution $\mathcal{D}$ et de fonction d'étiquetage f.



L'aire du grand carré vaut 2, celle du petit vaut 1. $\mathcal{D}$ distribue les exemples uniformément dans le grand carré et la fonction d'étiquetage f produit 1 si l'exemple (bleu) est dans le petit carré, sinon 0 (rouge).

Surapprentissage

Que donne la fonction suivante ?

$h_S(x) = y_i$ **si** $\exists i \in \{1, m\}$ tel que $x = x_i$
$h_S(x) = 1$ **sinon**

Elle produit une erreur apparente nulle (sur S).
Elle pourrait provenir d'un algorithme qui suit le principe ERM.

Son erreur *réelle* calculée sur $\mathcal{D}$ vaut 1/2 :

- Elle prédit 0 sur un nombre fini de cas.
- Elle prédit 1 sur un nombre infini, ce qui n'est vrai qu'une une fois sur deux.

Le modèle de génération des données

Hypothèse i.i.d.

On suppose que chaque x_i de S est indépendamment échantillonné selon $\mathcal{D}$; c'est l'hypothèse i.i.d. (indépendant et identiquement distribué).

Donc $S_x \sim \mathcal{D}^m$, où $S_x = \{x_1, ..., x_m\}$ est la représentation des exemples de S sur $\mathcal{X}$, et $\mathcal{D}^m$ est la distribution sur $\mathcal{X}^m$ obtenue par le produit de m distributions $\mathcal{D}$ sur $\mathcal{X}$.

Hypothèse de réalisabilité

On suppose qu'il existe f tel que : pour tout $x \in \mathcal{X}, y = f(x)$.

Plan

L'apprentissage PAC

- Étant donné que l'apprenant A a accès uniquement à S (et pas à $\mathcal{D}$ ni à f), il est illusoire que A puisse produire un h parfaitement correct, tel que $L_{\mathcal{D},f}(h) = 0$.
- On se contentera d'obtenir un h qui soit approximativement correct :

$$L_{\mathcal{D},f}(h) < \epsilon$$

- Étant donné qu'il est possible d'avoir un S peu représentatif de $\mathcal{D}$, on se contentera d'être probablement approximativement correct :

$$\mathbb{P}\Big(L_{\mathcal{D},f}(h) < \epsilon\Big) > 1 - \delta$$

c'est à dire correct avec une probabilité supérieure à $1 - \delta$ sur les tirages de S_x selon $\mathcal{D}^m$.

Le critère d'apprentissage PAC

Une classe $\mathcal{H}$ de classificateurs binaires est PAC-apprenable s'il existe une fonction $m_{\mathcal{H}} : [0,1]^2 \to \mathbb{N}$ et un algorithme d'apprentissage A tel que :

- pour tout ϵ et δ dans $[0,1]^2$
- pour toute distribution $\mathcal{D}$ sur $\mathcal{X}$,
- et pour toute cible $f \in \mathcal{H}$

l'apprenant A trouve une hypothèse h telle que :

$$\mathcal{D}^m\{S_x : L_{\mathcal{D},f}(h) \leq \epsilon\} \geq (1-\delta) \qquad \text{quand} \quad m \geq m_{\mathcal{H}}(\epsilon,\delta)$$

- $m_{\mathcal{H}}$ est appelée la complexité d'échantillon pour apprendre $\mathcal{H}$.
- Quelles sont alors les classes $\mathcal{H}$ « apprenables » au sens PAC ?

Un résultat négatif. . .mais pas tant que ça !

Théorème *No free lunch*

Pour tout apprenant A, il existe une tâche $(\mathcal{D}, f)$ sur laquelle A va échouer (au sens PAC) bien qu'il existe f d'erreur nulle.
Il n'existe donc pas d'algorithme d'apprentissage universel permettant d'apprendre $\mathcal{H}$ au sens PAC lorsque $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$.

Biais inductif

Il faut définir un biais d'apprentissage, c'est à dire se limiter à une famille d'hypothèses $\mathcal{H}$, soit finie, soit décrite par une formule.
Par exemple, dans $\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ pourrait être l'ensemble des rectangles et toute hypothèse serait forcément un rectangle.

Le principe ERM devient alors :

$$ERM_{\mathcal{H}}(S) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \quad L_S(h)$$

Plan

On va montrer que si $|\mathcal{H}|$ est fini, alors, pour $m = |S|$ assez grand, il n'y aura pas de surapprentissage.

Prendre les notes de cours.

Apprendre une classe finie

- Supposons que $\mathcal{H}$ est une classe comprenant un nombre fini d'hypothèses.
  - e.g. : $\mathcal{H}$ est l'ensemble des fonctions de $\mathcal{X}$ vers $\mathcal{Y}$ pouvant être implémentés par un programme d'au plus b bits ($|\mathcal{H}| = 2^{b+1} - 1$).
- Utilisons l'apprenant Cohérent :
  - Entrée : un échantillon $S = ((x_1, y_1), \ldots, (x_m, y_m))$.
  - Sortie : n'importe quel $h \in \mathcal{H}$ tel que $\forall i,\ y_i = h(x_i)$.
- C'est un cas particulier de l'algorithme de la Minimisation du Risque Empirique (ERM)

$\mathrm{ERM}_{\mathcal{H}}(S)$

- Entrée : l'échantillon $S = ((x_1, y_1), \ldots, (x_m, y_m))$.
- Définition du risque empirique : $L_S(h) = \frac{1}{m}|\{i : h(x_i) \neq y_i\}|$.
- Sortie : n'importe quel $h \in \mathcal{H}$ minimisant $L_S(h)$.

Apprendre une classe finie

Théorème

Soit $\mathcal{H}$ une classe finie de classificateurs binaires.

- *$\mathcal{H}$ est apprenable au sens PAC avec la complexité d'échantillon $m_{\mathcal{H}}(\epsilon, \delta) = \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$.*
- *Cette complexité d'échantillon est obtenue par l'algorithme de minimisation du risque empirique $\mathrm{ERM}_{\mathcal{H}}$.*

Pour démontrer ce théorème, il faut démontrer que pour tout $f \in \mathcal{H}$ et pour tout $\mathcal{D}$:

$$\mathcal{D}^m(\{S_x : L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) \leq \epsilon\}) \;\geq\; 1 - \delta\,.$$

Pourquoi cette formule ?

- Quelle est la performance de $ERM_{\mathcal{H}}$ sur un ensemble fini d'hypothèses ?
- $ERM_{\mathcal{H}}$ produit le résultat h_S avec $h_S = argmin_{h \in \mathcal{H}} L_s(h)$.
- Hypothèse de réalisabilité : Pout toute distribution $\mathcal{D}$ et toute fonction d'étiquetage f, il existe $h^* \in \mathcal{H}$ tel que $L_{\mathcal{D},f}(h^*) = 0$.
- Hypothèse i.i.d. Les éléments de S sont i.i.d. selon $\mathcal{D}$.
- Ce qui nous intéresse est le risque réel $L_{\mathcal{D},f}(h_S)$, et pas le risque apparent sur $L_S(h_S)$.

Le pourquoi de la formule

Puisque S est tiré aléatoirement i.i.d., $L_{\mathcal{D},f}(h_S)$ est une variable aléatoire.
$L_{\mathcal{D},f}(h_S) > \epsilon$ représente une erreur de l'apprentissage (approximativent) exact.
Il nous faut donc montrer que

$$\mathcal{D}^m(\{S_x : L_{\mathcal{D},f}(ERM_{\mathcal{H}}(S)) \leq \epsilon\}) > 1 - \delta$$

Qu'est-ce qu'une mauvaise hypothèse ?

h est mauvaise si $L_{\mathcal{D},f}(h) > \epsilon$
Les mauvaises hypothèses forment un sous-ensemble $\mathcal{H}_B$ (ou $\mathcal{H}_\epsilon$) de $\mathcal{H}$.

Qu'est-ce qu'un échantillon S trompeur ?

S est trompeur si $\exists h \in \mathcal{H}_B \ : \ L_S(h) = 0$
Les échantillons trompeurs forment un ensemble M.

Argument principal

L'hypothèse de réalisibilité implique que $L_S(h_S) = 0$. Donc, l'évènement $L_{\mathcal{D},f}(h_S) > \epsilon$ ne peut se produire que si pour une hypothèse h mauvaise on a $L_S(h) = 0$.
Autrement dit, l'évènement $L_{\mathcal{D},f}(h_S) > \epsilon$ ne peut se produire que si S est trompeur (élément de M).

Conclusion

$$\{S_x : L_{\mathcal{D},f}(h_S) > \epsilon\} \subseteq M$$

Autrement dit

$$M = \bigcup_{h \in \mathcal{H}_B} \{S_x : L_S(h) = 0\}$$

Finalement

$$\mathcal{D}^m\{S_x : L_{\mathcal{D},f}(h_S) > \epsilon\} \leq \mathcal{D}^m(M)$$

avec

$$\mathcal{D}^m(M) = \bigcup_{h \in \mathcal{H}_B} \{S_x : L_S(h)\} = 0$$

Union Bound Property : $P(A \cup B) \leq P(A) + p(B)$

$$\mathcal{D}^m(\{S_x : L_{\mathcal{D},f})(h_S) > \epsilon\} \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S_x : L_S(h) = 0\})$$

équation 2.8

- L'évènement $L_S(h) = 0$ est équivalent à l'évènement : $\forall i,\ h(x_i) = f(x_i)$
- Puisque S est tiré i.i.d.

$$\mathcal{D}^m(\{S_x : L_S(h) = 0\}) = \mathcal{D}^m(\{S_x : \forall i,\ h(x_i) = f(x_i)\})$$
$$= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\})$$

équation (2.8) du texte.

Pour chaque élément de S, puisque $h \in \mathcal{H}_D$, on a :

$$\mathcal{D}(\{x_i : h(x_i) = y_i\}) = 1 - L_{\mathcal{D},f}(h) \leq 1 - \epsilon$$

Résumé

$$\mathcal{D}^m(\{S_x : L_S(h) = 0\}) == \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\})$$

$$\mathcal{D}(\{x_i : h(x_i) = y_i\}) = 1 - L_{\mathcal{D},f}(h) \leq 1 - \epsilon$$

Rappel : $(1-\epsilon)^m \leq e^{-\epsilon m}$

$$\mathcal{D}^m(\{S_x : L_S(h) = 0\}) \leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

Conclusion

$$\mathcal{D}^m(\{S_x : L_S(h) > \epsilon\}) \leq |\mathcal{H}_B| e^{-\epsilon m} \leq |\mathcal{H}| e^{-\epsilon m}$$

Théorème PAC pour un espace d'hypothèses fini.

Soit un espace d'hypothèses $\mathcal{H}$ de taille finie.
Soit $\delta[0,1]$ et $\epsilon > 0$.
Soit m un entier tel que $m \geq \frac{log(|H|/\delta)}{\epsilon}$.
Alors, pour toute fonction d'étiquetage f et pour toute distribution $\mathcal{D}$ vérifiant l'hypothèse de réalisabilité,

- avec la probabilité au moins $1 - \delta$
- pour tout tirage i.i.d. d'un ensemble S de taille m
- pour toute hypothèse ERM h_S

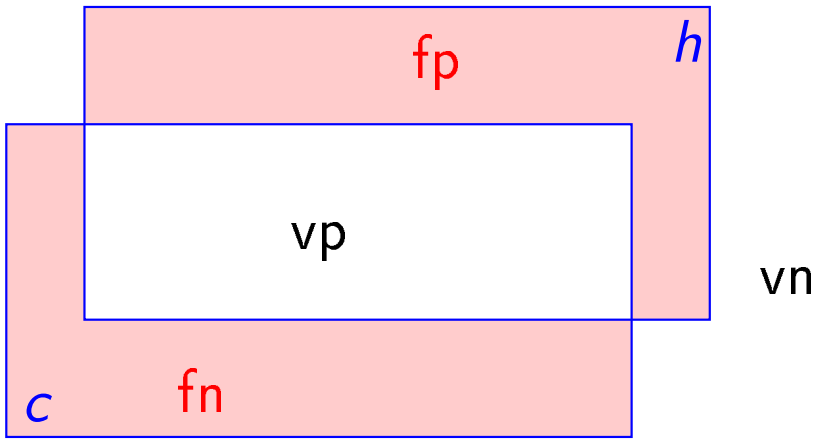
on a :

$$L_{\mathcal{D},f}(h_S) \leq \epsilon$$

Plan

Notations : exemples

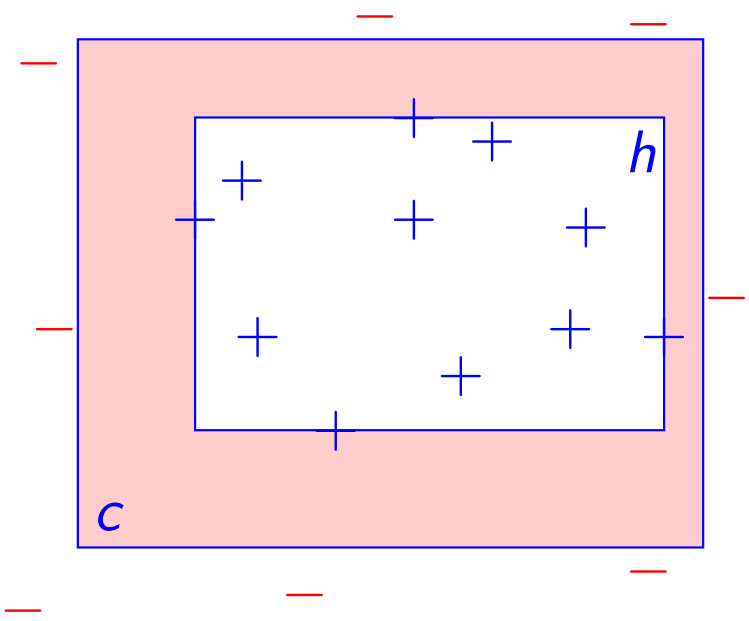
- $\mathcal{X}$: les points de $\mathbb{R}^2$ et $\mathcal{Y} = \{+1, -1\}$;
- $x = (70, 175)$;
- $\mathcal{H}$: les rectangles de $\mathbb{R}^2$ parallèles aux axes ;
- c un rectangle particulier :
  - vue ensembliste : les points contenus dans le rectangle c ;
  - vue fonctionnelle : test d'appartenance au rectangle c ;
- $\mathcal{D}$: distribution poids/taille chez l'homme ;
- erreur d'une hypothèse :



- Soit A l'algorithme qui retourne le plus petit rectangle h contenant tous les exemples négatifs de l'ensemble d'apprentissage (le moindre généralisé).
- L'algorithme A est ERM : il minimise les erreurs sur l'ensemble d'apprentissage.
- On suppose que le problème est réalisable, c'est à dire qu'il existe un rectangle cible c d'erreur réelle nulle.
- On va montrer que la classe infinie $\mathcal{H}$ des rectangles parallèles aux axes dans $\mathbb{R}^2$ est PAC-apprenable avec une complexité en échantillon valant

$$m_{\mathcal{H}} = \frac{4}{\epsilon} Log(\frac{4}{\delta})$$

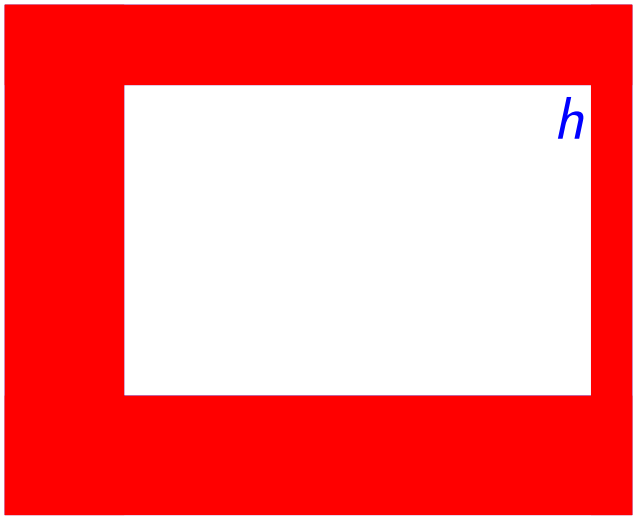
Démonstration (1) : c et h



Remarques

- h est toujours inclus dans c, car h moindre-généralisé ;
- h présente donc uniquement une erreur de type fn.

Démonstration (2) : borner l'erreur

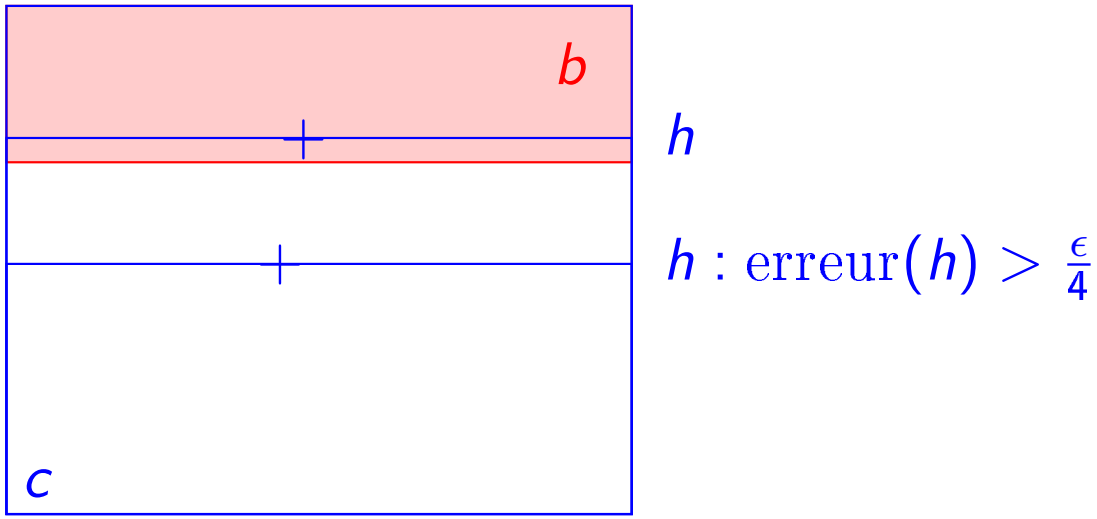


Objectifs

- limiter l'erreur globale à ϵ ;
- calculer la probabilité que l'une des bandes excède $\frac{\epsilon}{4}$.

Démonstration (3) : une bande

On définit une bande b de poids $\frac{\epsilon}{4}$ en haut de c.



Cas favorable : l'échantillon A contient un exemple (positif) dans b. L'erreur de h est inférieure à $\frac{\epsilon}{4}$ sur la bande haute.
Cas défavorable : $A \cap b = \emptyset$. L'erreur de h est supérieure à $\frac{\epsilon}{4}$ sur la bande haute. Probabilité de cette situation ?

Démonstration (4) : probabilités

- Tirer un exemple dans b : $\frac{\epsilon}{4}$;
- tirer un exemple en dehors de b : $1 - \frac{\epsilon}{4}$;
- tirer n exemples en dehors de b : $\left(1 - \frac{\epsilon}{4}\right)^n$;
- ne pas avoir d'exemple dans l'une des bandes : $\leq 4 \times \left(1 - \frac{\epsilon}{4}\right)^n$.

et on obtient finalement :

$$\Pr\left[\text{erreur}(h) > \epsilon\right] \leq 4 \times \left(1 - \frac{\epsilon}{4}\right)^n \leq 4 \times \left(e^{-\frac{\epsilon}{4}}\right)^n = 4 \times e^{-\frac{n\epsilon}{4}}$$

en utilisant le fait que : $(1 - x) \leq e^{-x}$.

Démonstration (5) : calcul du n minimal

On veut $\Pr[\text{erreur}(h) > \epsilon] \leq \delta$ et on résout donc :

$$\begin{array}{rrcl} & 4 \times e^{-\frac{n\epsilon}{4}} & \leq & \delta \\ \Rightarrow & e^{-\frac{n\epsilon}{4}} & \leq & \frac{\delta}{4} \\ \Rightarrow & -\frac{n\epsilon}{4} & \leq & \ln(\frac{\delta}{4}) \\ \Rightarrow & -n & \leq & \frac{4}{\epsilon} \times \ln(\frac{\delta}{4}) \\ \Rightarrow & n & \geq & -\frac{4}{\epsilon} \times \ln(\frac{\delta}{4}) \\ \Rightarrow & n & \geq & \frac{4}{\epsilon} \times \ln(\frac{4}{\delta}) \end{array}$$

Démonstration (6) : conclusions

- L doit constituer son échantillon avec $n = \frac{4}{\epsilon} \times \ln(\frac{4}{\delta})$;
- cela garantit : $\Pr[\text{erreur}(h) \leq \epsilon] \geq 1 - \delta$;
- L est linéaire en $\frac{1}{\epsilon}$;
- L est logarithmique en $\frac{1}{\delta}$.

La classe des rectangles parallèles aux axes dans $\mathbb{R}^2$ est efficacement PAC-apprenable.
(VC dimension de cette classe ?)

Plan

Le modèle PAC agnostique

- On a supposé que les étiquettes étaient générées par un $f \in \mathcal{H}$
- Cette supposition peut s'avérer trop forte !
- Maintenant, soyons plus réaliste en considérant que les étiquettes sont générées par une distribution (que nous ne connaissons pas).

PAC agnostique : définition formelle

Définition (Critère PAC agnostique)

Un classe d'hypothèses $\mathcal{H}$ est apprenable au sens PAC agnostique, relativement à un ensemble $Z = \mathcal{X} \times \mathcal{Y}$ et une une fonction de perte $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$, s'il existe une fonction $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ et un algorithme d'apprentissage A satisfaisant la propriété suivante : pour tout $\epsilon, \delta \in (0,1)$, $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ et distribution $\mathcal{D}$ sur Z, nous avons

$$\mathcal{D}^m \left(\left\{ S \in Z^m \ : \ L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right\} \right) \ \geq \ 1 - \delta$$

La propriété de convergence uniforme

Définition (convergence uniforme)

$\mathcal{H}$ possède la *propriété de convergence uniforme* (relativement à ℓ) s'il existe une fonction $m_{\mathcal{H}}^{\mathsf{UC}} : (0,1)^2 \to \mathbb{N}$ telle que pour tout $\epsilon, \delta \in (0,1)$, pour toute distribution $\mathcal{D}$, et pour tout $m \geq m_{\mathcal{H}}^{\mathsf{UC}}(\epsilon, \delta)$, nous avons

$$\mathcal{D}^m \left(\{ S \in Z^m : \ \forall h \in \mathcal{H}, \ \ |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \} \right) \geq 1 - \delta$$

Donc, avec probabilité $\geq 1 - \delta$, $L_S(h)$ est une bonne estimation de $L_{\mathcal{D}}(h)$ pour tout $h \in \mathcal{H}$ lorsque $\mathcal{H}$ satisfait la propriété de convergence uniforme.

La convergence uniforme suffit pour apprendre

Théorème (La convergence uniforme suffit pour apprendre)

- *Si $\mathcal{H}$ possède la propriété de convergence uniforme (relativement à ℓ) avec la fonction $m_{\mathcal{H}}^{UC}$, alors $\mathcal{H}$ est apprenable au sens PAC agnostique avec une complexité d'échantillon $m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$.*
- *Dans ce cas, $\mathrm{ERM}_{\mathcal{H}}$ est un algorithme d'apprentissage pour $\mathcal{H}$ au sens PAC agnostique.*

Preuve: Soit $h^* \in \mathrm{argmin}_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$ et $h_S \in \mathrm{argmin}_{h\in\mathcal{H}} L_S(h)$. Si $\mathcal{H}$ satisfait la propriété de convergence uniforme, alors lorsque $m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta)$, nous avons avec probabilité $\geq 1 - \delta$

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \tfrac{\epsilon}{2} \leq L_S(h^*) + \tfrac{\epsilon}{2} \leq L_{\mathcal{D}}(h^*) + \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = L_{\mathcal{D}}(h^*) + \epsilon\,.$$

Puisque $h_S = \mathrm{ERM}_{\mathcal{H}}(S)$ et $L_{\mathcal{D}}(h^*) = \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$, on a

$$\mathcal{D}^m\left(\left\{S \in Z^m :\ L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) \ \leq\ \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\right\}\right) \geq 1 - \delta$$

Les classes finies sont apprenables au sens PAC agnostique

Nous allons démontrer le théorème suivant :

Théorème

Soit $\mathcal{H}$ une classe finie et soit une fonction de perte à valeur dans $[0,1]$. Alors, $\mathcal{H}$ est apprenable au sens PAC agnostique en utilisant $\mathrm{ERM}_{\mathcal{H}}$ avec la complexité d'échantillon satisfaisant

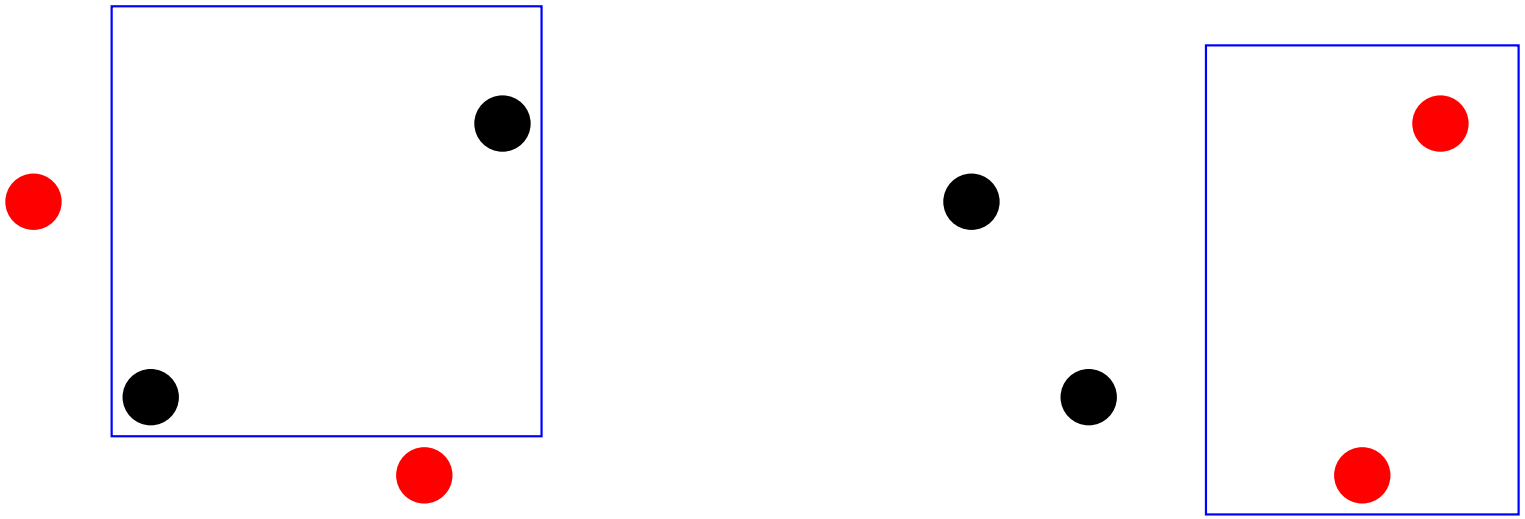
$$m_{\mathcal{H}}(\epsilon,\delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$

Preuve: En raison du dernier théorème, il suffit de démontrer que $\mathcal{H}$ possède la propriété de convergence uniforme avec

$$m^{\mathrm{UC}}_{\mathcal{H}}(\epsilon,\delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil .$$

Plan

Pulvérisation (*shattering*)



Est-ce qu'un rectangle (côtés parallèles aux axes) peut toujours séparer quatre points du plan selon deux classes
quelque soit l'étiquetage de ces quatre points ?

Pulvérisation d'un ensemble C étiqueté en deux classes par une famille d'hypothèses $\mathcal{H}$

- Soit $C = \{x_1, ..., x_{|C|}\} \subset \mathcal{X}$.
- $\mathcal{H}_C$ dénote la restriction de $\mathcal{H}$ sur C, autrement dit :

$$\mathcal{H}_C = \{(h(x_1), \cdots, h(x_{|C|})) \in \{\pm 1\}^{|C|} : h \in \mathcal{H}\}$$

- Alors, $|\mathcal{H}_C| \leq 2^{|C|}$.
- On dit que $\mathcal{H}$ pulvérise C si $|\mathcal{H}_C| = 2^{|C|}$.

VC-dimension

Par définition :

$$VCdim(H) = sup\{|C| : \mathcal{H} \text{ pulvérise } C\}$$

VC-dimension : calcul

Soit un ensemble d'hypothèses $\mathcal{H}$. Pour montrer que $VCdim(H) = d$ il faut prouver deux choses :

1. Il existe un ensemble C de taille d qui est pulvérisé par $\mathcal{H}$.
2. Aucun ensemble de taille $d+1$ ne peut être pulvérisé par $\mathcal{H}$.

Attention !

Pulvériser un ensemble C signifie : séparer les deux classes pour tous les étiquetages possibles.

VC-dimension : exemples

Les rectangles parallèles aux axes dans $\mathbb{R}^2$

- On sait que $VCdim(\mathcal{H}_{rect}) \geq 4$, puisqu'un ensemble C de 4 points du plan peut toujours être séparé en ses deux classes par un rectangle.
- Il est facile de voir que ce n'est plus le cas pour $|C| = 5$.
- Donc $VCdim(\mathcal{H}_{rect}) = 4$

Les droites dans $\mathbb{R}^2$

$VCdim(\mathcal{H}_{droites}) = 3$

Un cas intéressant

$\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$ avec $h_\theta : \mathcal{X} \to \{-1, 1\}$ défini par $h_\theta = \lceil 0.5 \sin(\theta x) \rceil$.

$$VCdim(\mathcal{H}) = \infty$$

Quelque soit d, il existe un certain ensemble C de taille d pulvérisé par une certaine hypothèse $h_\theta \in \mathcal{H}$.

La dimension VC — exemples

Classes finies :

- Soit $d =$ la dimension VC d'un $\mathcal{H}$ fini.
- Or pour pulvériser d points, il faut au moins 2^d fonctions.
- Donc, $|\mathcal{H}| \geq 2^d$, car d points sont pulvérisés.
- Donc, $d \leq \log_2(|\mathcal{H}|)$. i.e., $\mathrm{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.
- Cependant il existe $\mathcal{H}$ ayant $\mathrm{VCdim}(\mathcal{H}) \ll \log_2(|\mathcal{H}|)$. e.g., les souches de décision avec seuil $\theta \in \mathbb{R}$ constituent une classe contenant une infinité de classificateurs mais $\mathrm{VCdim}(\mathcal{H}) = 1$ pour cette classe.

Plan

Le théorème fondamental de l'apprentissage statistique

Soit $\mathcal{H}$ une famille de classificateurs binaires telle que $VCdim(\mathcal{H}) = d$.
Il existe alors deux constantes positives c_1 et c_2 telles que la complexité d'échantillon $m_{\mathcal{H}}(\epsilon, \delta)$ nécessaire pour apprendre $\mathcal{H}$ au sens PAC satisfait :

$$c_1 \frac{d \log(1/\epsilon)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq c_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

Cette complexité d'échantillon $m_{\mathcal{H}}(\epsilon, \delta)$ peut être obtenue par un algorithme ERM de minimisation du risque empirique.