IFT-7002 Fondements de l'apprentissage machine

L'apprentissage probablement approximativement correct (PAC)

Shai Shalev-Shwartz
The Hebrew University of Jerusalem

Traduit et adapté par Mario Marchand
Université Laval

Hiver 2024

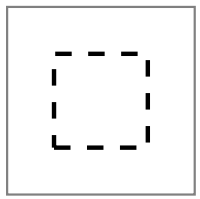
Sommaire

Définitions

- Le domaine, $\mathcal{X}$: L'ensemble des *instances* (*i.e.*, objets) possibles que l'on désire étiqueter.
- L'ensemble des étiquettes, $\mathcal{Y}$.
- Un prédicteur, $h : \mathcal{X} \to \mathcal{Y}$: utilisé pour étiqueter les instances. Cette fonction est également appelée une *hypothèse*, ou un *classificateur* (lorsque $\mathcal{Y}$ est discret).

Exemple :

- $\mathcal{X} = \mathbb{R}^2$ représentant la couleur et la dureté des papayes.
- $\mathcal{Y} = \{\pm 1\}$ représentant "savoureux" ou "non-savoureux".
- $h(x) = 1$ si x est dans le rectangle interne



Apprentissage **en lots ("Batch")**

- L'entrée de l'apprenant A (*i.e.*, algorithme d'apprentissage) :
  - Échantillon d'apprentissage S de m *exemples*,
    $S \stackrel{\text{def}}{=} ((x_1, y_1) \dots (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$
    - Chaque exemple est une paire (instance, étiquette).
- La sortie $A(S)$ de l'apprenant A :
  - Une règle de prédiction, $h : \mathcal{X} \to \mathcal{Y}$. Donc, $h = A(S)$.
- Quel devrait être l'objectif de l'apprenant ?
- Intuitivement, $A(S)$ devrait être correct sur les exemples à venir.

“Correct sur les exemples à venir”

- Considérons donc (dans ce chapitre) que les étiquettes sont produites par un classificateur f que nous appelons la **cible**.
- Définissons alors l'*erreur* (ou *risque*) de h relativement à la cible f par

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)]$$

où $\mathcal{D}$ est une **distribution** (inconnue) sur $\mathcal{X}$. i.e., pour tout $A \subset \mathcal{X}$, la valeur de $\mathcal{D}(A)$ est la probabilité que $x \in A$.

- Alors,

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

- Le but de A est alors de trouver h tel que $L_{\mathcal{D},f}(h)$ soit petit.

Le modèle de génération des données

- Puisque $h = A(S)$, nous devons alors supposer qu'il existe une relation entre les données S d'apprentissage et $(\mathcal{D}, f)$.
- Pour cela, on suppose que **chaque x_i de S est indépendamment échantillonné selon $\mathcal{D}$** ; ce que l'on dénote par l'hypothèse i.i.d. (indépendant et identiquement distribué).
  - Donc $S_x \sim \mathcal{D}^m$, où $S_x \overset{\text{def}}{=} (x_1, \ldots, x_m)$ est la projection 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 $\mathcal{X}$.
- Réalisable : Il existe f tel que : pour tout $x \in \mathcal{X}$, $y = f(x)$.
- On généralisera au cas non réalisable dès le prochain chapitre.
- De plus, on se limite au cas de la classification binaire ($\mathcal{Y} = \{\pm 1\}$) dans ce chapitre. On généralisera dès le prochain chapitre.

Probablement Approximativement Correct (PAC)

- Étant donné que A a accès uniquement à S (et non à $\mathcal{D}$ et f), il est illusoire que A puisse produire h tel que $L_{\mathcal{D},f}(h) = 0$
- On se contentera d'obtenir h qui soit approximativement correct :
  - *i.e.*, $L_{\mathcal{D},f}(h) \leq \epsilon$.
- Étant donné qu'il est possible d'obtenir S qui soit peu représentatif de $\mathcal{D}$, on se contentera d'être probablement approximativement correct :
  - *i.e.*, $L_{\mathcal{D},f}(A(S)) \leq \epsilon$ avec probabilité $\geq 1 - \delta$ sur les tirages de $S_x \sim \mathcal{D}^m$.

Apprentissage PAC

Définition (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 $\epsilon, \delta \in (0,1)$,
- pour toute distribution $\mathcal{D}$ sur $\mathcal{X}$, et pour toute cible $f \in \mathcal{H}$,

nous avons

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

lorsque $m \geq m_{\mathcal{H}}(\epsilon, \delta)$.

$m_{\mathcal{H}}$ est appelé la complexité d'échantillon pour apprendre $\mathcal{H}$.

Quelles sont alors les classes $\mathcal{H}$ "apprenables" au sens PAC ?

Apprentissage PAC

Leslie Valiant, gagnant du prix Turing 2010

> *"For transformative contributions to the theory of computation, including the theory of probably approximately correct (PAC) learning, the complexity of enumeration and of algebraic computation, and the theory of parallel and distributed computing".*

Le théorème "No Free Lunch"

Théorème ("No Free Lunch")

- *Considérons un algorithme d'apprentissage A qui reçoit un échantillon S tel que $|S| < |\mathcal{X}|/2$ (et tel que $y \in \{-1, +1\}$ pour tout $(x, y) \in S$).*
- *Alors, il existe $\mathcal{D}, f$ tel que $L_{\mathcal{D},f}(A(S)) \geq 1/8$ avec probabilité $\geq 1/7$.*

- Donc, 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 $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$.
- Idée de la preuve (voir sect 5.1 pour la preuve exacte) :
  - Soit un ensemble fini $C \subset \mathcal{X}$ et une distribution $\mathcal{D}$ uniforme sur C.
  - Si $m \leq |C|/2$, A n'a aucune information sur les étiquettes d'au moins la moitié des éléments de C. Il existe alors **plusieurs** f contredisant les étiquette prédites par $A(S)$ sur cette moitié non observée.

Connaissance a priori (biais d'apprentissage)

- La classe $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ n'est donc pas "apprenable" au sens PAC.
- Quelles sont alors les classes $\mathcal{H}$ "apprenables" au sens PAC ?
- L'apprenant A doit donc se servir de l'information que $f \in \mathcal{H}$.
- C'est un biais d'apprentissage que nous fournissons à A.
- Plus généralement, le théorème précédent démontre la nécessité d'avoir un biais (ou une connaissance a priori) pour pouvoir apprendre.

Sommaire

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 $\mathcal{D}$:

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

Preuve

- De manière équivalente, il faut démontrer que $\forall\ f \in \mathcal{H},\ \forall\ \mathcal{D}$, on a :

$$\mathcal{D}^m(\{S_x : L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \leq \delta$$

- Soit $\mathcal{H}_\epsilon$ l'ensemble des hypothèses ϵ-mauvaises,

$$\mathcal{H}_\epsilon \stackrel{\text{def}}{=} \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}$$

- Soit M, l'ensemble des échantillons trompeurs ("misleading"),

$$M \stackrel{\text{def}}{=} \{S_x : \exists h \in \mathcal{H}_\epsilon, L_S(h) = 0\} = \bigcup_{h \in \mathcal{H}_\epsilon} \{S_x : L_S(h) = 0\}$$

- Observons que :

$$\{S_x : L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\} \subseteq M = \bigcup_{h \in \mathcal{H}_\epsilon} \{S_x : L_S(h) = 0\}$$

Preuve (suite)

Lemme (Borne de l'union)

Pour tout ensembles A, B et distribution $\mathcal{D}$, nous avons que

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B) \ .$$

- Nous avons démontré que :
  $\{S_x : L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\} \subseteq \bigcup_{h \in \mathcal{H}_\epsilon} \{S_x : L_S(h) = 0\}$
- Alors, par la borne de l'union, nous avons que

$$\begin{aligned}
\mathcal{D}^m(\{S_x : & L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\}) \\
& \leq \sum_{h \in \mathcal{H}_\epsilon} \mathcal{D}^m(\{S_x : L_S(h) = 0\}) \\
& \leq |\mathcal{H}_\epsilon| \max_{h \in \mathcal{H}_\epsilon} \mathcal{D}^m(\{S_x : L_S(h) = 0\})
\end{aligned}$$

Preuve (suite)

- Observons que pour tout $f \in \mathcal{H}$ et pour tout $h \in \mathcal{H}$, on a :

$$\mathcal{D}^m(\{S_x : L_S(h) = 0\}) \;=\; (1 - L_{\mathcal{D},f}(h))^m$$

- Si $h \in \mathcal{H}_\epsilon$, alors $L_{\mathcal{D},f}(h) > \epsilon$. Alors

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

- Nous avons alors démontré que pour tout $f \in \mathcal{H}$, on a :

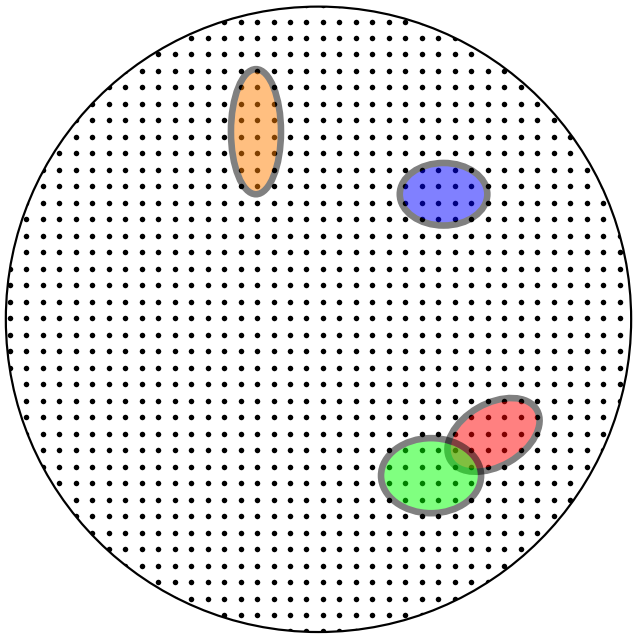
$$\mathcal{D}^m(\{S_x : L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) \geq \epsilon\}) \;<\; |\mathcal{H}_\epsilon|\,(1 - \epsilon)^m$$

- Finalement, en utilisant $1 - \epsilon \leq e^{-\epsilon}$ et $|\mathcal{H}_\epsilon| \leq |\mathcal{H}|$ nous concluons que

$$\mathcal{D}^m(\{S_x : L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) \geq \epsilon\}) \;<\; |\mathcal{H}|\,e^{-\epsilon\, m}$$

- Le terme de droite est $\leq \delta$ lorsque $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$. □

Illustration de l'utilisation de la borne de l'union



- Chaque point représente un échantillon possible S. Chaque ellipse couvre les échantillons trompeurs pour un $h \in \mathcal{H}_\epsilon$. La probabilité de chaque ellipse est $\leq (1-\epsilon)^m$. L'algorithme retournera possiblement un $h \in \mathcal{H}_\epsilon$ si l'échantillon tombe dans l'un de ces ellipses.

Sommaire

Qu'est-ce qui est PAC-apprenable et quels sont les algorithmes pour y arriver ?

- Que se passe-t-il pour les classes infinies d'hypothèses ?
- Quelle est la complexité d'échantillon pour une classe donnée ?
- Existe-t-il un algorithme générique possédant une complexité d'échantillon optimale ?

Qu'est-ce qui est PAC-apprenable et quels sont les algorithmes pour y arriver ?

Le théorème fondamental de l'apprentissage :

- La complexité d'échantillon de $\mathcal{H}$ est caractérisée par la dimension VC de $\mathcal{H}$, *i.e.*, $\mathrm{VCdim}(\mathcal{H})$.
- L'algorithme $\mathrm{ERM}_{\mathcal{H}}$ est un algorithme générique pour apprendre toute classe $\mathcal{H}$ dont la $\mathrm{VCdim}(\mathcal{H})$ est finie.

Chervonenkis

Vapnik

La dimension VC — définition exacte

- Soit $C = \{x_1, \ldots, x_{|C|}\} \subset \mathcal{X}$.
- $\mathcal{H}_C$ dénote la restriction de $\mathcal{H}$ sur C, i.e.,

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

- Alors, $|\mathcal{H}_C| \leq 2^{|C|}$.
- Nous disons que $\mathcal{H}$ pulvérise C si $|\mathcal{H}_C| = 2^{|C|}$.
- $\mathrm{VCdim}(\mathcal{H}) \stackrel{\text{def}}{=} \sup\{|C| \ : \ \mathcal{H} \text{ pulvérise } C\}$.
- Lorsque $C \subset \mathcal{X}$ est pulvérisé par $\mathcal{H}$, les étiquettes d'un sous ensemble de C ne procurent pas assez d'information sur la cible f.

Pour trouver $\mathrm{VCdim}(\mathcal{H})$

Pour démontrer que $\mathrm{VCdim}(\mathcal{H}) = d$, nous devons démontrer que :

1. Il existe un ensemble C de taille d qui est pulvérisé par $\mathcal{H}$.
2. Il n'existe pas d'ensemble C de taille $d+1$ qui est pulvérisé par $\mathcal{H}$.

La dimension VC — exemples

Les souches de décision : $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathbb{R}\}$.
Convention utilisée tout au long du cours : pour tout $a \in \mathbb{R}$, on a

$$\text{sign}(a) \;=\; \begin{cases} 1 & \text{si} \quad a > 0 \\ -1 & \text{si} \quad a \leq 0\,. \end{cases}$$

- Notez que $\{0\}$ est pulvérisé
- Mais aucun ensemble de 2 points $\{x_1, x_2\}$ (avec $x_1 < x_2$ s.p.d.g.) ne peut être pulvérisé car il est impossible d'assigner $+1$ à x_1 et -1 à x_2 par une souche de décision.
- Donc $\text{VCdim}(\mathcal{H}) = 1$ pour les souches de décision.

La dimension VC — exemples

Les intervalles : $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_{a,b} : a < b \in \mathbb{R}\}$, où $h_{a,b}(x) = 1$ ssi $x \in [a, b]$

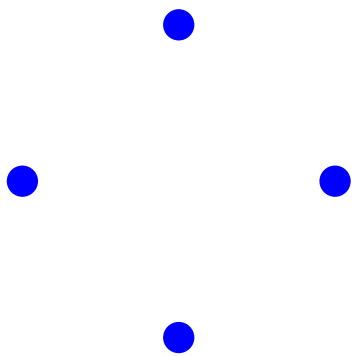
- Notez que que $\{0, 1\}$ est pulvérisé
- Mais aucun triplet de points $\{x_1, x_2\, x_3\}$ (avec $x_1 < x_2 < x_3$ s.p.d.g.) ne peut être pulvérisé car il est impossible d'assigner $+1$ à x_1 et x_3 et d'assigner -1 à x_2 à l'aide d'un intervalle sur $\mathbb{R}$.
- Donc $\mathrm{VCdim}(\mathcal{H}) = 2$ pour les intervalles sur $\mathbb{R}$.

La dimension VC — exemples

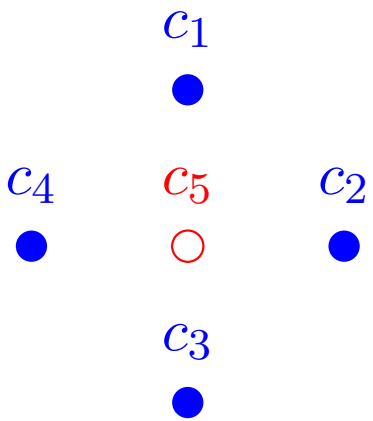
Les rectangles à axes parallèles : $\mathcal{X} = \mathbb{R}^2$,
$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 < a_2 \text{ et } b_1 < b_2\}$, où $h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = 1$ ssi $x_1 \in [a_1, a_2]$ et $x_2 \in [b_1, b_2]$

Il est possible de pulvériser 4 points comme c'est le cas à gauche, mais aucun ensemble de 5 points ne peut être pulvérisé.

Pulvérisé



Non pulvérisé



Alors $\mathrm{VCdim}(\mathcal{H}) = 4$ pour les rectangles à axes parallèles sur $\mathbb{R}^2$.

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.

La dimension VC des demi-espaces

Demi-espaces homogènes : $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H}_h^d = \{\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}$

- Soit $\mathbf{e}_i$ un "one hot vector", i.e., toute les composantes de $\mathbf{e}_i$ sont nulles sauf sa i-ième composante qui est $= 1$.
- $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ est pulvérisé car pour avoir $\langle \mathbf{w}, \mathbf{e}_i \rangle = y_i$ pour tout $i \in [d]$ quelque soit $y_1, \ldots, y_d$, il suffit de choisir $w_i = y_i$ pour tout i.
- Maintenant, soit un ensemble $\{\mathbf{x}_1, \ldots, \mathbf{x}_{d+1}\}$ de $d+1$ points de $\mathbb{R}^d$.
- Alors il existe $a_1, \ldots, a_{d+1}$ (pas tous nuls) tel que $\sum_{i=1}^{d+1} a_i \mathbf{x}_i = \mathbf{0}$.
- Soit $I \stackrel{\text{def}}{=} \{i : a_i > 0\}$ et $J \stackrel{\text{def}}{=} \{i : a_i < 0\}$.
- On a, s.p.d.g., que I est non vide puisque $\sum_{i=1}^{d+1} (-a_i) \mathbf{x}_i = \mathbf{0}$.
- Par contre, il est possible que J soit vide. De plus on a

$$\sum_{i \in I} a_i \mathbf{x}_i = \sum_{i \in J} |a_i| \mathbf{x}_i \,.$$

- Donc, chacune des sommes donne $\mathbf{0}$ lorsque J est vide.

La dimension VC des demi-espaces

- Si $\{\mathbf{x}_1, \ldots, \mathbf{x}_{d+1}\}$ est pulvérisé, il existe $\mathbf{w}$ tel que $\langle \mathbf{w}, \mathbf{x}_i \rangle > 0$ pour tout $i \in I$ et $\langle \mathbf{w}, \mathbf{x}_i \rangle \leq 0$ pour tout $i \in J$.
- Si J est non vide, on a la contradiction :

$$0 < \sum_{i \in I} a_i \langle \mathbf{x}_i, \mathbf{w} \rangle = \langle \sum_{i \in I} a_i \mathbf{x}_i, \mathbf{w} \rangle = \langle \sum_{i \in J} |a_i| \mathbf{x}_i, \mathbf{w} \rangle = \sum_{i \in J} |a_i| \langle \mathbf{x}_i, \mathbf{w} \rangle \leq 0$$

- Si J est vide, nous obtenons aussi une contradiction :

$$0 < \sum_{i \in I} a_i \langle \mathbf{x}_i, \mathbf{w} \rangle = \langle \sum_{i \in I} a_i \mathbf{x}_i, \mathbf{w} \rangle = \langle \mathbf{0}, \mathbf{w} \rangle = 0 \,.$$

- On obtient alors toujours une contradiction si $\{\mathbf{x}_1, \ldots, \mathbf{x}_{d+1}\}$ est pulvérisé.
- Donc, aucun ensemble de $d+1$ points ne peut être pulvérisé par les demi-espaces homogènes dans $\mathbb{R}^d$. Donc $\mathrm{VCdim}(\mathcal{H}_h^d) = d$.

La dimension VC des demi-espaces

Demi-espaces NON homogènes :

$\mathcal{H}_{nh}^d = \{\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b) : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$

- Notez que $\{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_d\}$ est pulvérisé par $\mathcal{H}_{nh}^d$ car pour tout $(y_1, \ldots, y_{d+1})$, avec $y_i = \pm 1$, on a $\exists \mathbf{w}, b$:

$$\langle \mathbf{w}, \mathbf{e}_i \rangle + b = y_i \quad \text{pour } i = 1, \ldots, d$$
$$\langle \mathbf{w}, \mathbf{0} \rangle + b = y_{d+1} \,.$$

- Supposons que $\{\mathbf{x}_1, \ldots, \mathbf{x}_{d+2}\}$ est pulvérisé par $\mathcal{H}_{nh}^d$.
- Il est possible de décrire les demi-espaces non homogènes par des demi-espace homogènes en ajoutant une coordonnée fixée à 1 à toutes les instances $\mathbf{x}$ et en ajoutant une composante b à $\mathbf{w}$ pour obtenir $\mathbf{x}' \stackrel{\text{def}}{=} (1, \mathbf{x})$ et $\mathbf{w}' \stackrel{\text{def}}{=} (b, \mathbf{w})$. Donc, $\langle \mathbf{w}', \mathbf{x}' \rangle = \langle \mathbf{w}, \mathbf{x} \rangle + b$.
- Dans ce cas, $\{(1, \mathbf{x}_1), \ldots, (1, \mathbf{x}_{d+2})\}$ est pulvérisé par $\mathcal{H}_h^{d+1}$.
- Puisque cela contredit le résultat précédent ($\text{VCdim}(\mathcal{H}_h^{d+1}) = d + 1$), aucun ensemble de $d + 2$ points n'est pulvérisé par $\mathcal{H}_{nh}^d$.
- Alors, $\text{VCdim}(\mathcal{H}_{nh}^d) = d + 1$.

Le théorème fondamental de l'apprentissage statistique

Théorème (Le théorème fondamental de l'apprentissage statistique)

Soit $\mathcal{H}$ une classe de classificateurs binaires dont $\mathrm{VCdim}(\mathcal{H}) = d$. Il existe alors des constantes positives C_1, C_2 telle que la complexité d'échantillon $m_{\mathcal{H}}$ pour apprendre $\mathcal{H}$ au sens PAC satisfait

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

De plus, cette complexité d'échantillon $m_{\mathcal{H}}$ est obtenue par l'algorithme $\mathrm{ERM}_{\mathcal{H}}$ (minimisation du risque empirique).

- Steve Haneke (JMLR, 2016) a obtenu la meilleure borne supérieure possible pour un algorithme différent de $\mathrm{ERM}_{\mathcal{H}}$:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq C_3 \frac{d + \log(1/\delta)}{\epsilon} .$$

Preuve de la borne supérieure – étapes initiales

- Considérons $\mathcal{H}$, tel que $\mathrm{VCdim}(\mathcal{H}) = d$. Il faut démontrer que pour tout $f \in \mathcal{H}$ et pour tout $\mathcal{D}$:

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

- De manière équivalente, il faut démontrer que $\forall\ f \in \mathcal{H},\ \forall\ \mathcal{D}$, on a :

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

- Soit $\mathcal{H}_\epsilon \stackrel{\text{def}}{=} \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}$.
- Or, $\{S_x : L_{\mathcal{D},f}(\mathrm{ERM}_{\mathcal{H}}(S)) > \epsilon\} \subseteq \{S_x : \exists h \in \mathcal{H}_\epsilon : L_S(h) = 0\}$.
- Il est donc suffisant de démontrer que $\forall\ f \in \mathcal{H},\ \forall\ \mathcal{D}$, on a :

$$\mathcal{D}^m(\{S_x : \exists h \in \mathcal{H}_\epsilon : L_S(h) = 0\}) \leq \delta\,,$$

- que nous pouvons ré-écrire

$$\mathop{\mathbb{P}}_{S_x \sim \mathcal{D}^m}[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0] \leq \delta\,.$$

Preuve de la borne supérieure – première étape

- Si $\mathcal{H}$ est infini, nous ne pouvons donc pas utiliser la borne de l'union directement comme dans le cas fini.
- Mais nous pourrons l'utiliser après l'introduction d'un échantillon additionnel fictif.

Théorème (Borne du double échantillon)

Lorsque $m \geq \frac{8\ln(2)}{\epsilon}$, *nous avons*

$$\begin{aligned}&\mathop{\mathbb{P}}_{S_x \sim \mathcal{D}^m}[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0] \\ &\qquad \leq 2 \mathop{\mathbb{P}}_{S_x, T_x \sim \mathcal{D}^m}[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2]\,.\end{aligned}$$

Notez que la condition $m \geq \frac{8\ln(2)}{\epsilon}$ est satisfaite par la borne supérieure du théorème fondamental pour $C_2 = 8\ln(2)$.

Preuve de la borne du double échantillon

- Utilisons $\mathbb{1}_{[a]} = 1$ si a est vrai et $\mathbb{1}_{[a]} = 0$ si a est faux.

$$\begin{aligned}
&\mathop{\mathbb{P}}_{S_x, T_x \sim \mathcal{D}^m}[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2] \\
&= \mathop{\mathbb{E}}_{S_x, T_x \sim \mathcal{D}^m} \mathbb{1}_{[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2]} \\
&= \mathop{\mathbb{E}}_{S_x, T_x \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}_\epsilon} \mathbb{1}_{[L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2]} \\
&= \mathop{\mathbb{E}}_{S_x, T_x \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}_\epsilon} \left(\mathbb{1}_{[L_S(h) = 0]} \, \mathbb{1}_{[L_T(h) \geq \epsilon/2]} \right) \\
&\geq \mathop{\mathbb{E}}_{S_x \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}_\epsilon} \left(\mathbb{1}_{[L_S(h) = 0]} \mathop{\mathbb{E}}_{T_x \sim \mathcal{D}^m} \mathbb{1}_{[L_T(h) \geq \epsilon/2]} \right) \\
&= \mathop{\mathbb{E}}_{S_x \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}_\epsilon} \left(\mathbb{1}_{[L_S(h) = 0]} \mathop{\mathbb{P}}_{T_x \sim \mathcal{D}^m}[L_T(h) \geq \epsilon/2] \right) \\
&\geq \mathop{\mathbb{E}}_{S_x \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}_\epsilon} \left(\mathbb{1}_{[L_S(h) = 0]} \mathop{\mathbb{P}}_{T_x \sim \mathcal{D}^m}[L_T(h) \geq L_{\mathcal{D},f}(h)/2] \right)
\end{aligned}$$

Car $L_{\mathcal{D},f}(h) > \epsilon \; \forall h \in \mathcal{H}_\epsilon$.

Preuve de la borne du double échantillon

- Utilisons une borne multiplicative de Chernoff pour borner $\mathbb{P}_{T_x \sim \mathcal{D}^m}[L_T(h) \geq L_{\mathcal{D},f}(h)/2]$.

Lemme (Bornes multiplicatives de Chernoff)

Soit $Z_1, \ldots, Z_m$ tels que $\mathbb{P}[Z_i = 1] = p$ et $\mathbb{P}[Z_i = 0] = 1 - p$. Soit $\hat{Z} \stackrel{def}{=} \frac{1}{m} \sum_{i=1}^{m} Z_i$. Pour tout $\gamma > 0$, nous avons

$$\mathbb{P}[\hat{Z} \leq (1-\gamma)p] \leq e^{-\frac{\gamma^2 pm}{2}}$$

$$\mathbb{P}[\hat{Z} \geq (1+\gamma)p] \leq e^{-\frac{\gamma^2 pm}{3}} .$$

La première inégalité pour $Z_i = \mathbb{1}_{[h(x_i) \neq f(x_i)]}$, $\hat{Z} = L_T(h)$, $p = L_{\mathcal{D},f}(h)$ et $\gamma = 1/2$, nous donne

$$\mathbb{P}_{T_x \sim \mathcal{D}^m}[L_T(h) \geq L_{\mathcal{D},f}(h)/2] \;\geq\; 1 - \exp\left(-\frac{L_{\mathcal{D},f}(h)m}{8}\right) \geq 1/2$$

lorsque $L_{\mathcal{D},f}(h) \geq \frac{8\ln(2)}{m}$.

Preuve de la borne du double échantillon

Donc, lorsque $L_{\mathcal{D},f}(h) \geq \frac{8\ln(2)}{m}$, nous avons

$$\begin{aligned}
&\mathop{\mathbb{P}}_{S_x,T_x\sim\mathcal{D}^m}[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2] \\
&\geq \mathop{\mathbb{E}}_{S_x\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}_\epsilon} \left(\mathbb{1}_{[L_S(h)=0]} \mathop{\mathbb{P}}_{T_x\sim\mathcal{D}^m}[L_T(h) \geq L_{\mathcal{D},f}(h)/2] \right) \\
&\geq \frac{1}{2} \mathop{\mathbb{E}}_{S_x\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}_\epsilon} \mathbb{1}_{[L_S(h)=0]} \\
&= \frac{1}{2} \mathop{\mathbb{P}}_{S_x\sim\mathcal{D}^m}[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0]
\end{aligned}$$

Ce qui donne le théorème de la borne du double échantillon pour $L_{\mathcal{D},f}(h) > \epsilon \geq \frac{8\ln(2)}{m}$, $\forall h \in \mathcal{H}_\epsilon$. □

Preuve de la borne supérieure

Donc, en utilisant la borne du double échantillon, nous avons

$$
\begin{aligned}
&\mathop{\mathbb{P}}_{S_x \sim \mathcal{D}^m} [\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0] \\
&\quad \leq 2 \mathop{\mathbb{P}}_{S_x, T_x \sim \mathcal{D}^m} [\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2] \\
&\quad \leq 2 \mathop{\mathbb{P}}_{S_x, T_x \sim \mathcal{D}^m} [\exists h \in \mathcal{H} : L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2] \\
&\quad = 2 \mathop{\mathbb{P}}_{S_x, T_x \sim \mathcal{D}^m} [\exists h \in \mathcal{H}_{S_x \cup T_x} : L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2] \\
&\quad = 2 \mathop{\mathbb{E}}_{S_x, T_x \sim \mathcal{D}^m} \mathbb{1}_{[\exists h \in \mathcal{H}_{S_x \cup T_x} : L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2]} \\
&\quad = 2 \mathop{\mathbb{E}}_{S_x, T_x \sim \mathcal{D}^m} \max_{h \in \mathcal{H}_{S_x \cup T_x}} \mathbb{1}_{[L_S(h) = 0 \text{ et } L_T(h) \geq \epsilon/2]} \\
&\quad = 2 \mathop{\mathbb{E}}_{S_x, T_x \sim \mathcal{D}^m} \max_{h \in \mathcal{H}_{S_x \cup T_x}} \mathbb{1}_{[L_S(h) = 0]} \mathbb{1}_{[L_T(h) \geq \epsilon/2]}
\end{aligned}
$$

où $\mathcal{H}_{S_x \cup T_x}$ désigne la restriction de $\mathcal{H}$ sur $S_x \cup T_x$.

Borne supérieure – symmétrisation

$$\mathop{\mathbb{P}}_{S_x\sim\mathcal{D}^m}[\exists h\in\mathcal{H}_\epsilon : L_S(h)=0] \leq 2 \mathop{\mathbb{E}}_{S_x,T_x\sim\mathcal{D}^m} \max_{h\in\mathcal{H}_{S_x\cup T_x}} \mathbb{1}_{[L_S(h)=0]}\mathbb{1}_{[L_T(h)\geq\epsilon/2]}$$

$$\leq 2 \mathop{\mathbb{E}}_{S_x,T_x\sim\mathcal{D}^m} \sum_{h\in\mathcal{H}_{S_x\cup T_x}} \mathbb{1}_{[L_S(h)=0]}\mathbb{1}_{[L_T(h)\geq\epsilon/2]}$$

- Symmétrisation : Puisque S, T sont i.i.d., nous pouvons considérer que l'on tire d'abord $2m$ exemples et, qu'ensuite, nous formons S en choisissant, au hasard uniforme, les exemples dans $S\cup T \stackrel{\text{def}}{=} A$.
- Considérons $J \stackrel{\text{def}}{=} \{\mathbf{j}\subset[2m] : |\mathbf{j}|=m\}$, l'ensemble des vecteurs $\mathbf{j}$ de m indices distincs. Notons par $A_\mathbf{j}$ le sous ensemble des m exemples de A pointés par $\mathbf{j}$ et $\mathbf{j}_S \stackrel{\text{def}}{=} (1,\ldots,m)$. Alors :

$$\mathop{\mathbb{E}}_{S_x,T_x\sim\mathcal{D}^m} \sum_{h\in\mathcal{H}_{S_x\cup T_x}} \mathbb{1}_{[L_S(h)=0]}\mathbb{1}_{[L_T(h)\geq\epsilon/2]} = \mathop{\mathbb{E}}_{A_x\sim\mathcal{D}^{2m}} \sum_{h\in\mathcal{H}_{A_x}} \mathbb{1}_{[L_{A_{\mathbf{j}_S}}(h)=0]}\mathbb{1}_{[L_A(h)\geq\epsilon/4]}$$

$$= \mathop{\mathbb{E}}_{A_x\sim\mathcal{D}^{2m}} \sum_{h\in\mathcal{H}_{A_x}} \mathbb{1}_{[L_A(h)\geq\epsilon/4]}\mathbb{1}_{[L_{A_\mathbf{j}}(h)=0]} \quad \forall\mathbf{j}\in J$$

Borne supérieure – symmétrisation

car pour toute fonction $f(A, A_{\mathbf{j}})$, la valeur de $\mathbb{E}_{A_x \sim \mathcal{D}^{2m}} f(A, A_{\mathbf{j}})$ est la même pour tout $\mathbf{j} \in J$.

- Soit $U(J)$ la distribution sur J telle que chaque composante j_k de $\mathbf{j}$ est distribuée uniformément sur $\{1, \ldots, 2m\} \setminus \{j_1, \ldots, j_{k-1}\}$. Alors, $\forall \mathbf{j} \in J$, on a

$$\begin{aligned}
&\mathop{\mathbb{E}}_{A_x \sim \mathcal{D}^{2m}} \sum_{h \in \mathcal{H}_{A_x}} \mathbb{1}_{[L_A(h) \geq \epsilon/4]} \mathbb{1}_{[L_{A_{\mathbf{j}}}(h) = 0]} \\
&= \mathop{\mathbb{E}}_{\mathbf{j} \sim U(J)} \mathop{\mathbb{E}}_{A_x \sim \mathcal{D}^{2m}} \sum_{h \in \mathcal{H}_{A_x}} \mathbb{1}_{[L_A(h) \geq \epsilon/4]} \mathbb{1}_{[L_{A_{\mathbf{j}}}(h) = 0]} \\
&= \mathop{\mathbb{E}}_{A_x \sim \mathcal{D}^{2m}} \sum_{h \in \mathcal{H}_{A_x}} \mathbb{1}_{[L_A(h) \geq \epsilon/4]} \mathop{\mathbb{E}}_{\mathbf{j} \sim U(J)} \mathbb{1}_{[L_{A_{\mathbf{j}}}(h) = 0]}
\end{aligned}$$

- On a donc

$$\mathop{\mathbb{P}}_{S_x \sim \mathcal{D}^m}[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0] \leq 2 \mathop{\mathbb{E}}_{A_x \sim \mathcal{D}^{2m}} \sum_{h \in \mathcal{H}_{A_x}} \mathbb{1}_{[L_A(h) \geq \epsilon/4]} \mathop{\mathbb{E}}_{\mathbf{j} \sim U(J)} \mathbb{1}_{[L_{A_{\mathbf{j}}}(h) = 0]}$$

Borne supérieure – symmétrisation

- La dernière espérance est la probabilité que h n'effectue aucune erreur sur les m exemples choisis dans A lorsque h fait $2m\epsilon/4$ erreurs sur A.
- Cette probabilité est alors au plus $(1-\epsilon/4)^m \leq e^{-\epsilon m/4}$. Alors :

$$\begin{aligned}
&\mathop{\mathbb{E}}_{A_x\sim\mathcal{D}^{2m}} \sum_{h\in\mathcal{H}_{A_x}} \mathbb{1}_{[L_A(h)\geq\epsilon/4]} \mathop{\mathbb{E}}_{\mathbf{j}\sim U(J)} \mathbb{1}_{[L_{A_\mathbf{j}}(h)=0]} \\
&\leq \mathop{\mathbb{E}}_{A_x\sim\mathcal{D}^{2m}} \sum_{h\in\mathcal{H}_{A_x}} \mathbb{1}_{[L_A(h)\geq\epsilon/4]}\ e^{-\epsilon m/4} \\
&\leq \mathop{\mathbb{E}}_{A_x\sim\mathcal{D}^{2m}} \sum_{h\in\mathcal{H}_{A_x}} e^{-\epsilon m/4} \ \leq \ \mathop{\mathbb{E}}_{A_x\sim\mathcal{D}^{2m}} |\mathcal{H}_{A_x}| e^{-\epsilon m/4} \\
&\leq \max_{C\subset\mathcal{X}:|C|=2m} |\mathcal{H}_C| e^{-\epsilon m/4}
\end{aligned}$$

Ainsi, nous avons obtenu que

$$\mathop{\mathbb{P}}_{S_x\sim\mathcal{D}^m}[\exists h\in\mathcal{H}_\epsilon : L_S(h)=0] \ \leq \ 2 \max_{C\subset\mathcal{X}:|C|=2m} |\mathcal{H}_C| e^{-\epsilon m/4}\,.$$

Fonction de croissance ("Growth function")

La dernière équation fait intervenir la **fonction de croissance** $\tau_{\mathcal{H}}$ sur $2m$ instances d'une classe $\mathcal{H}$ de classificateurs. Par définition, nous avons

$$\tau_{\mathcal{H}}(m) \stackrel{\text{def}}{=} \max_{C \subset \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

Ce qui donne le nombre maximal de dichotomies qu'il est possible de réaliser sur m points à l'aide des fonctions de $\mathcal{H}$.

Lemme (Sauer-Shelah-Perles-Vapnik-Chervonenkis)

Pour toute classe $\mathcal{H}$ de classificateurs tel que $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$, et pour tout $0 < d \leq m$, nous avons

$$\tau_{\mathcal{H}}(m) \;\leq\; \sum_{i=0}^{d} \binom{m}{i} \;\leq\; \left(\frac{em}{d}\right)^{d}.$$

Voir les notes de Maria-Florina Balcan pour la preuve.

Borne supérieure – étape finale

Donc, bien que $\tau_{\mathcal{H}}(m) = 2^m$ lorsque $m \leq d = \mathrm{VCdim}(\mathcal{H})$, $\tau_{\mathcal{H}}(m)$ a une croissance **polynomiale en** m lorsque $m > d$. Si nous appliquons ce lemme à notre borne supérieure, nous avons alors

$$\mathop{\mathbb{P}}_{S_x \sim \mathcal{D}^m}[\exists h \in \mathcal{H}_\epsilon : L_S(h) = 0] \;\leq\; 2 \cdot \left(\frac{2em}{d}\right)^d \cdot e^{-\epsilon m/4}$$

La borne supérieure est alors obtenue en bornant supérieurement le terme à droite par δ. On cherche alors m pour avoir

$$e^{\epsilon m/4} \;\geq\; \frac{2}{\delta}\left(\frac{2em}{d}\right)^d,$$

ou, de manière équivalente, on cherche m pour satisfaire

$$m \;\geq\; \frac{4}{\epsilon}\left[d\ln(2em/d) + \ln(2/\delta)\right] \;=\; \frac{4d}{\epsilon}\ln(m) + \frac{4}{\epsilon}\left[d\ln(2e/d) + \ln(2/\delta)\right].$$

Borne supérieure – étape finale

Or, selon le lemme A.2 du manuel, pour tout $a \geq 1$ et pour tout $b \geq 0$:

$$x \geq 4a \ln(2a) + 2b \implies x \geq a \ln(x) + b \,.$$

Donc, pour avoir

$$m \geq \frac{4d}{\epsilon} \ln(m) + \frac{4}{\epsilon} \left[d \ln(2e/d) + \ln(2/\delta) \right] ,$$

il suffit d'avoir

$$m \geq \frac{16d}{\epsilon} \ln\left(\frac{8d}{\epsilon}\right) + \frac{8}{\epsilon} \left[d \ln(2e/d) + \ln(2/\delta) \right] .$$

Pour cela, il suffit d'avoir

$$\begin{aligned} m &\geq \frac{16d}{\epsilon} \ln\left(\frac{8d}{\epsilon}\right) + \frac{16}{\epsilon} \left[d \ln(2e/d) + \frac{1}{2} \ln(2/\delta) \right] \\ &= \frac{16d}{\epsilon} \ln\left(\frac{16e}{\epsilon}\right) + \frac{8}{\epsilon} \ln(2/\delta) \,. \end{aligned}$$

Ce qui est conforme à la borne supérieure du théorème fondamental. □

Sommaire

Minimisation du risque empirique pour les demi-espaces

- Rappel :
$$\mathcal{H} = \{\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}$$
- $\text{ERM}_{\mathcal{H}}$ pour demi-espaces :
Soit $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$.
Trouver $\mathbf{w}$ tel que pour tout i : $\text{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle) = y_i$ (avec $y_i \in \{\pm 1\}$).
- Ceci est un programme linéaire :
Trouver $\mathbf{w}$ tel que $\forall i, \quad y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0$.
- Peut être résolu efficacement à l'aide de la programmation linéaire
  - Algorithme de Karmarkar (en temps polynomial)
  - Méthode du simplex (non polynomial en pire cas)
- Peut être résolu par l'algorithme du Perceptron (lorsqu'il existe une solution).

$\mathrm{ERM}_{\mathcal{H}}$ à l'aide de l'algorithme du perceptron

Algorithme du perceptron

initialiser : $\mathbf{w} = (0, \ldots, 0) \in \mathbb{R}^d$
Tant que $\exists i$ tel que $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \leq 0$
$\quad \mathbf{w} \leftarrow \mathbf{w} + y_i \mathbf{x}_i$

- Proposé par Rosenblatt en 1958.

Le théorème de convergence du perceptron

Théorème (Agmon'54, Block'62, Novikoff'62)

Soit $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ *un ensemble d'exemples tel qu'il existe* $\mathbf{w}^* \in \mathbb{R}^d$ *possédant la propriété que pour tout* i, $y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq 1$. *Alors, le perceptron fera au plus*

$$\|\mathbf{w}^*\|^2 \max_i \|\mathbf{x}_i\|^2$$

mises à jour pour obtenir un demi-espace cohérent avec les exemples.

- La condition s'applique ssi il existe un demi-espace cohérent avec les exemples. On dit alors que les données sont **linéairement séparables**.
- Par contre, $\|\mathbf{w}^*\|$ pourrait devoir être très grand afin de satisfaire $y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq 1 \ \forall i$. Ce qui nécessiterait beaucoup de mises à jour.
- Dans plusieurs cas, il est possible que $\|\mathbf{w}^*\|$ ne soit pas trop grand.

Preuve

- Soit $(\mathbf{x}_t, y_t)$ l'exemple utilisé pour la t-ième mise à jour de $\mathbf{w}$
- Soit $\mathbf{w}^{(t)}$ la valeur de $\mathbf{w}$ juste avant la mise à jour t. Donc

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_t \mathbf{x}_t, \text{ lorsque } y_t \langle \mathbf{w}^{(t)}, \mathbf{x}_t \rangle \leq 0$$

- Soit $R = \max_i \|\mathbf{x}_i\|$
- Le cosinus de l'angle entre $\mathbf{w}^*$ et $\mathbf{w}^{(t)}$ est donné par $\frac{\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle}{\|\mathbf{w}^{(t)}\| \, \|\mathbf{w}^*\|}$
- Or, le cosinus doit être toujours ≤ 1 (inégalité de Cauchy-Schwarz)
- Nous allons démontrer que :
  1. $\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle \geq t$
  2. $\|\mathbf{w}^{(t+1)}\| \leq R\sqrt{t}$
- Ceci impliquera alors que :

$$\frac{t}{R\sqrt{t}\,\|\mathbf{w}^*\|} \leq \frac{\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle}{\|\mathbf{w}^{(t+1)}\| \, \|\mathbf{w}^*\|} \leq 1$$

- Ce qui implique $t \leq \|\mathbf{w}^*\|^2 R^2$ (le résultat désiré).

Preuve (suite)

Démonstration que $\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle \geq t$:

- Initialement, $\langle \mathbf{w}^{(1)}, \mathbf{w}^* \rangle = 0$ car $\mathbf{w}^{(1)} = \mathbf{0}$.
- Lors d'une mise à jour, $\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle$ augmente par au moins 1 :

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle = \langle \mathbf{w}^{(t)} + y_t \mathbf{x}_t, \mathbf{w}^* \rangle = \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle + \underbrace{y_t \langle \mathbf{x}_t, \mathbf{w}^* \rangle}_{\geq 1}$$

Démonstration que $\|\mathbf{w}^{(t+1)}\|^2 \leq R^2\, t$:

- Initialement, $\|\mathbf{w}^{(1)}\|^2 = 0$ car $\mathbf{w}^{(1)} = \mathbf{0}$.
- Lors d'une mise à jour, $\|\mathbf{w}^{(t)}\|^2$ augmente par au plus R^2 :

$$\begin{aligned} \|\mathbf{w}^{(t+1)}\|^2 &= \|\mathbf{w}^{(t)} + y_t \mathbf{x}_t\|^2 = \|\mathbf{w}^{(t)}\|^2 + \underbrace{2 y_t \langle \mathbf{w}^{(t)}, \mathbf{x}_t \rangle}_{\leq 0} + y_t^2 \|\mathbf{x}_t\|^2 \\ &\leq \|\mathbf{w}^{(t)}\|^2 + R^2 \ . \end{aligned}$$

Résumé

- Le modèle d'apprentissage PAC
- Qu'est-il possible d'apprendre au sens PAC ?
- Apprentissage PAC des classes finies par $\mathrm{ERM}_{\mathcal{H}}$
- La dimension VC et le théorème fondamental de l'apprentissage statistique
- Les classes dont la dimension VC est finie sont apprenables au sens PAC par $\mathrm{ERM}_{\mathcal{H}}$
- Apprendre efficacement les demi-espaces avec la programmation linéaire et l'algorithme du perceptron lorsque les données sont linéairement séparables