# Apprentissage (Machine learning)

Notes par Sarah Vita Paardekooper

08 Octobre 2020

## 3 Agnostic setting

Soient $X$ un domaine, $Y$ un ensemble d'étiquettes et $\mathcal{H} \subset \{f : X \to Y\}$ un ensemble d'hypothèses.

Dans le modèle de cohérence (consistency model), on suppose qu'il existe un $c \in \mathcal{H}$ tel que l'échantillon $S$ est cohérent avec $c$, c'est à dire $S = \{(x_i, c(x_i))\}_i$. L'objectif est donc de trouver une hypothèse $h \in \mathcal{H}$ cohérent avec $S$.

Dans le modèle PAC, on suppose qu'il existe un $c \in \mathcal{H}$ tel que l'échantillon $S$ est cohérent avec $c$, c'est à dire $S = \{(x_i, c(x_i))\}_i$. Mais on suppose que les $x_i$ peuvent contenir du bruit, donc on suppose une distribution $\mathcal{D}$ sur le domaine $X$ et $S$ suit $\mathcal{D}^m$. L'objectif est donc de trouver une hypothèse $h \in \mathcal{H}$ tel que $\mathbb{P}_{S\sim\mathcal{D}^m}(\mathrm{err}_c(h) \leq \epsilon) \geq 1 - \delta$ où $\mathrm{err}_c(h) = \mathbb{P}_{X\sim\mathcal{D}}(h(x) = c(x))$.

Dans le modèle PAC agnostic, on suppose l'existence de bruit sur le domaine ainsi que sur les étiquettes, c'est à dire une distribution $\mathcal{D}$ sur $X \times Y$. L'échantillon $S = \{(x_i, y_i)\}_i$ suit alors une distribution $\mathcal{D}^m$. Dans ce modèle, l'objectif est de trouver $h \in \mathcal{H}$ tel que

$$\mathbb{P}_{S\sim\mathcal{D}^m}(\mathrm{err}(h) - \inf_{h'\in\mathcal{H}} \mathrm{err}(h') \leq \epsilon) \geq 1 - \delta$$

où $\mathrm{err}(h) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}(h(x) \neq y)$.

L'**erreur empirique** est définit par $\widehat{\mathrm{err}}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h(x_i)\neq y_i}$, pour $h \in \mathcal{H}$.

**Définition 3.1.** On dit que $h_{ERM} \in \mathcal{H}$ est une hypothèse **ERM** (Empirical Risk Minimisation) si $\widehat{\mathrm{err}}(h_{ERM}) = \min_{h\in\mathcal{H}} \widehat{\mathrm{err}}(h)$, c'est à dire si

$$h_{ERM} = \arg\min_{h\in\mathcal{H}} \widehat{\mathrm{err}}(h).$$

Un algorithme est **ERM** s'il retourne une hypothèse ERM.

**Définition 3.2.** On dit que $\mathcal{H}$ est **apprenable au sens agnostic PAC** s'il existe un algorithme calculant une hypothèse $h$ tel que $\forall\epsilon, \forall\delta, \forall\mathcal{D}$ sur $X \times Y, \exists M \in \mathbb{N}, \forall m \geq M$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}(\mathrm{err}(h) - \inf_{h'\in\mathcal{H}} \mathrm{err}(h') \leq \epsilon) \geq 1 - \delta$$

**Théorème 3.3.** *$\mathcal{H}$ est apprenable au sens agnostic PAC si et seulement si $\mathcal{H}$ a VC-dimension finie.*

La VC-dimension est une mesure de la capacité d'apprendre un ensemble de fonction par classification statistique binaire.

**Définition 3.4.** La Propriété de Convergence Uniforme (**UCP**) est définit comme suit :

$$\forall h \in \mathcal{H}, \mathbb{P}_{(X,Y)\sim\mathcal{D}}(h(x) \neq y) = |\operatorname{err}(h) - \widehat{\operatorname{err}}(h)| \leq \frac{\epsilon}{2}$$

**Lemme 3.5.** *Si UCP est vraie (pour $\epsilon$ fixéé), alors l'algorithme ERM ayant $h_{ERM}$ en sortie satisfait $\operatorname{err}(h_{ERM}) - \inf_{h\in\mathcal{H}} \operatorname{err}(h) \leq \epsilon$.*

*Preuve.* Notons $h_{OPT} = \arg\inf_{h\in\mathcal{H}} \operatorname{err}(h)$. Montrons que $\operatorname{err}(h_{ERM}) - \operatorname{err}(h_{OPT}) \leq \epsilon$ :

$$\begin{aligned}
\operatorname{err}(h_{ERM}) - \operatorname{err}(h_{OPT}) &\leq \underbrace{\operatorname{err}(h_{ERM}) - \widehat{\operatorname{err}}(h_{ERM})}_{\leq \epsilon/2 \text{ (UCP)}} + \underbrace{\widehat{\operatorname{err}}(h_{ERM})}_{\widehat{\operatorname{err}}(h_{OPT}) \text{ (ERM)}} - \operatorname{err}(h_{OPT}) \\
&\leq \tfrac{\epsilon}{2} + \underbrace{\widehat{\operatorname{err}}(h_{OPT}) - \operatorname{err}(h_{OPT})}_{\leq \epsilon/2 \text{ (UCP)}} \\
&\leq \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon
\end{aligned}$$

□

Soient un domaine $X$ et un ensemble d'étiquettes $Y = \{0, 1\}$. Notons que l'ensemble $\mathcal{H}$ de fonctions $X \to \{0, 1\}$ est équivalent à l'ensemble de sous-ensembles de $X$ (puisque toutes les fonctions peuvent être exprimées sous la forme $\mathbb{1}_A$, avec $A \subseteq X$).

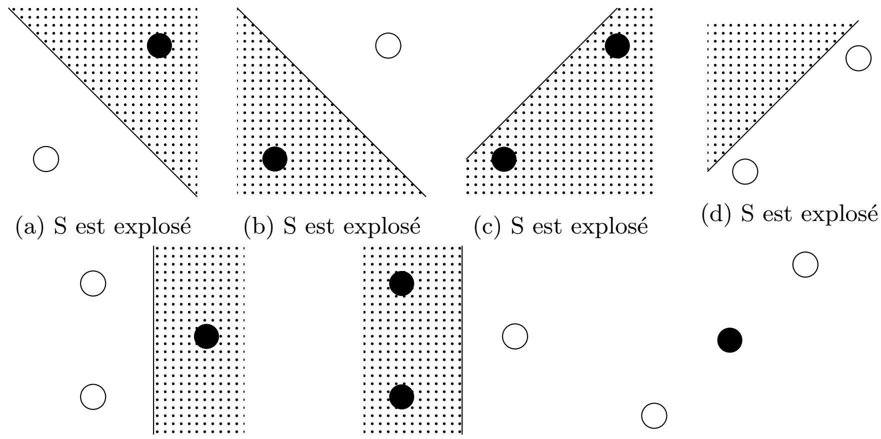
**Définition 3.6.** Soit $S \subseteq X$ un échantillon. On dit que $S$ est **explosé** (shattered) par $\mathcal{H}$ si $\forall P \subseteq S, \exists h \in \mathcal{H}, \forall x \in S, h(x) = 1 \Leftrightarrow x \in P$.

Exemple :
$X = \mathbb{R}^n$, $\mathcal{H}_n = \{f : \mathbb{R}^n \to \{0,1\} : f(\overrightarrow{x}) = 1 \text{ si } \overrightarrow{x} \cdot \overrightarrow{a} \geq b \text{ et } 0 \text{ si } \overrightarrow{x} \cdot \overrightarrow{a} < b\}$, avec $\overrightarrow{a} \in \mathbb{R}^n$ et $b \in \mathbb{R}$. $\mathcal{H}_n$ est l'ensemble des fonctions dans $\mathbb{R}^n$ définient par un hyperplan, avec tout les points valant 1 d'une part de l'hyperplan, et 0 de l'autre.

- n=2 :
  - Si $|S| = 2$, $S$ est toujours explosé.
  - Il existe des échantillons $S$, $|S| = 3$ explosés et il existent des échantillons $S$, $|S| = 3$ non explosés.
  - Tout échantillon $S$, $|S| \geq 4$, est non explosé

**Définition 3.7.** On définit alors la **VC-dimension** comme

$$\operatorname{VCdim}(\mathcal{H}) = \max\{|S| : S \text{ explosé par } \mathcal{H}\}$$



(a) S est explosé (b) S est explosé (c) S est explosé (d) S est explosé

(e) S est explosé (f) S est explosé (g) S n'est pas explosé

Figure 1: Les points noirs sont dans $P$, et la zone en pointillés vaut 1.

Exemples : $\mathrm{VCdim}(\mathcal{H}_2) = 3$, $\mathrm{VCdim}(\mathcal{H}_n) = n+1$. $X = \mathbb{N}$, $\mathcal{H}$ est l'ensemble de toutes les fonctions, $\mathrm{VCdim}(\mathcal{H}) = \infty$.

**Définition 3.8.** La **fonction de croissance** de $\mathcal{H}$ est définit par

$$\begin{aligned} \Pi_{\mathcal{H}} :& \mathbb{N} \to \mathbb{N} \\ & m \mapsto \max_{S \subseteq X, |S| = m} |\{h|_S : h \in \mathcal{H}\}| \end{aligned}$$

*Remarque* 3.9. Pour $\mathrm{VCdim}(\mathcal{H}) = d$, on a $\Pi_{\mathcal{H}}(d+1) < 2^{d+1}$, $\Pi_{\mathcal{H}}(d) = 2^d$, et $\Pi_{\mathcal{H}}(d') = 2^{d'}$ pour tout $d' \leq d$.

On a aussi que $S$ est explosé par $\mathcal{H}$ si et seulement si $|\{h|_S : h \in \mathcal{H}\}| = 2^d$.

**Lemme 3.10.** *(Sauer) Si VCdim($\mathcal{H}$) = $d$, alors*

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