Cours 2 - PAC : Probably Approximately Correct

Notes par Adrien Maurin

Définition du problème Soient X un ensemble de données (« data points ») et Y (équivalent à $\{0,1\}$) un ensemble de labels. Étant donné un couple $S = \{(x_0, y_0), \cdots, (x_m, y_m)\}$, l'objectif est de trouver $h : X \to Y$ avec $h \in \mathcal{H}$ tel que h trouve au mieux S.

Définition 1 (Consistancy model)**.** *La fonction h du problème défini précédemment vérifie la condition suivante :* $\forall i \in [1, m], h(x_i) = y_i$.

Remarque

1. La fonction n'assure pas la généralité.
2. La fonction ne prend pas en compte le bruit.

Définition 2 (PAC model)**.** *Ce modèle repose sur l'hypothèse suivante : on suppose que les données sont générées par une distribution $\mathcal{D}$ inconnue tel que :* $\mathcal{D} : X \longrightarrow [0,1]$ *et* $\sum_{x \in X} \mathcal{D}(x) = 1$.

On va introduire deux paramètres :
— ϵ : la précision
— δ : la confiance

Algorithme d'apprentissage Un algorithme d'apprentissage fonctionne de la manière suivante :
Input : $S = \{(x_0, y_0), \cdots, (x_m, y_m)\}$
Output : $h \in \mathcal{H}$.

Définition 3 (erreur)**.** *On défini l'erreur d'une fonction (notée err) de la manière suivante :*

$$\forall \epsilon > 0, \forall \delta > 0, \forall c \in \mathcal{H} \text{ (objectif) }, \forall \mathcal{D}, \exists M \in \mathbb{N}, \forall m \geqslant M :$$
$$err(h) = \mathbb{P}_{x\tilde{}\mathcal{D}}(h(x) \neq c(x))$$

La notation $x\tilde{}\mathcal{D}$ signifie que l'on tire un élément x avec la distribution $\mathcal{D}$.

Définition 4 (apprentissage)**.** *Un algorithme apprend au sens PAC la classe $\mathcal{H}$ si :*

$$\boxed{\mathbb{P}_{x\tilde{}\mathcal{D}^m}(err(h) \leqslant \epsilon) \geqslant 1 - \delta} \tag{1}$$

En des termes plus simples, cela signifie qu'avec une forte probabilité, mon erreur est faible. On rajoute l'adjectif efficacement si $m = poly(\frac{1}{\epsilon}; \frac{1}{\delta})$.

Exemple 1 : ensemble de points sur $\mathbb{R}^2$

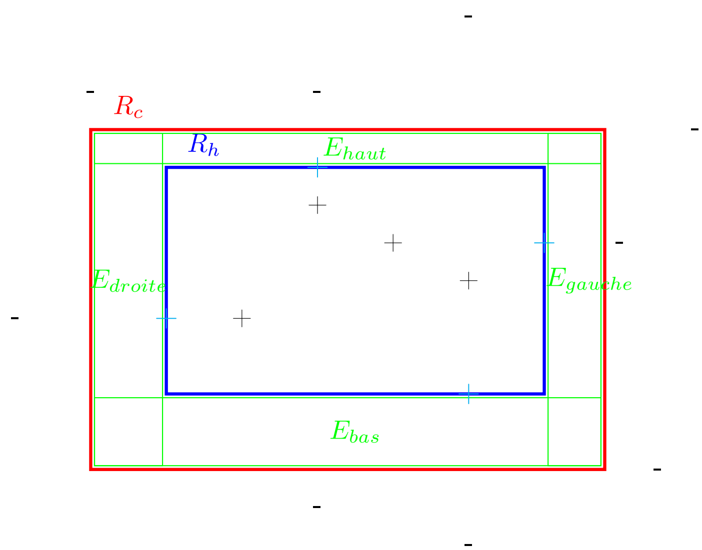

FIGURE 1 – Illustration du problème.

Description du problème Soit un plan sur $\mathbb{R}^2$. Sur ce plan, on a des symboles $-$ et des $+$. L'objectif est de déterminer le plus grand rectangle tel que les symboles $+$ sont dans le rectangle. On définit R_h le plus petit rectangle qui contient les $+$. Le rectangle R_c étant celui qu'on cherche à déterminer.

On cherche à déterminer $\mathbb{P}_{x\check{}\mathcal{D}^m}(\text{err}(h) \leqslant \epsilon) \geqslant 1-\delta$. Pour cela, on pose : $\mathbb{P}_{x\check{}\mathcal{D}}(E_{gauche}) = \mathbb{P}_{x\check{}\mathcal{D}} = \frac{\epsilon}{4}$. On fait de même pour E_{droite}, E_{bas} et E_{haut}. On définit l'évènement A suivant :

$$A = \{\exists x, y, z, t \in S, x \in E_{bas} \wedge y \in E_{haut} \wedge z \in E_{droite} \wedge t \in E_{gauche}\}.$$

Si A est réalisé, alors on a $\text{err}_c(h) \leqslant \epsilon$. En effet :

$$\text{err}_c(h) = \mathbb{P}_{x\check{}\mathcal{D}}(h(x) \neq c(x)) = \mathbb{P}_{x\check{}\mathcal{D}}(x \in R_c \backslash R_h) = \mathbb{P}(R_c \backslash R_h)$$

Si l'évènement A est réalisé, on obtient $R_c \backslash R_h \subseteq E_{gauche} \cup E_{droite} \cup E_{bas} \cup E_{haut}$. On a alors l'inégalité suivante :

$$\mathbb{P}(R_c \backslash R_h) \leqslant \mathbb{P}(E_{gauche} \cup E_{droite} \cup E_{bas} \cup E_{haut}) \leqslant \sum \mathbb{P}(E_i) = \epsilon$$

On a le résultat suivant : $\mathbb{P}(\forall x \in S, x \notin E_{bas}) = (1 - \frac{\epsilon}{4})^m$. On en déduit donc par le complémentaire que :

$$\mathbb{P}(\exists x \in S, x \in E_{bas}) = 1 - (1 - \frac{\epsilon}{4})^m \leqslant 1 - e^{-\frac{m\epsilon}{4}}$$

En revenant à l'évènement A, on obtient : $\mathbb{P} \geqslant 1 - 4e^{-\frac{m\epsilon}{4}}$.

$$\mathbb{P}(A^C) = \mathbb{P}(\forall x \in S, x \notin E_{bas} \cup \cdots \cup x \notin E_{gauche}) \leqslant \sum \mathbb{P}(\forall x \in S, x \notin E_i) \leqslant 4e^{-\frac{m\epsilon}{4}}$$

On en déduit le résultat : $\mathbb{P}(A) \geqslant 1 - 4e^{-\frac{m\epsilon}{4}}$. En combinant ce résultat avec l'équation (1), il faut choisir δ tel que $\delta \leqslant 4e^{-\frac{m\epsilon}{4}}$. Cette inégalité est vraie si $\boxed{m \geqslant -\frac{4}{\epsilon} \ln(\frac{\delta}{4})}$.

Exemple 2 : formules booléennes conjonctives

Soient $X = \{0,1\}^n$, $Y = \{0,1\}$ et $\mathcal{H} = \{\phi = \bigwedge_i l_i : l_i = p_i \text{ ou } \neg p_i\}$.
Par exemple avec $n = 4$, $\phi = p_1 \wedge p_3 \wedge \neg p_4$

Algorithme :

— On commence avec $\phi = p_1 \wedge \neg p_1 \wedge \cdots \wedge p_n \wedge \neg p_n$.
— Pour chaque $x \in S$ positif,on enlève les littéraux incompatiples dans ϕ.

Objectif
L'objectif est de trouver $m \in \mathbb{N}$ qui vérifie l'équation (1). On défini la fonction erreur par : $\text{err}(h) = \mathbb{P}_{x \tilde{} \mathcal{D}}(x \models \phi_c \not\Leftrightarrow \phi_h) \leqslant \epsilon$

Définition 5 (mauvais littéral)**.** *Soit l un littéral. On dit qu'un littéral est mauvais si $\mathbb{P}_{x \tilde{} \mathcal{D}}(x \models \phi_c \wedge \phi \not\models l) \geqslant \frac{\epsilon}{2n}$*

Propriété 1. *Si ϕ_h ne contient aucun mauvais littéral, alors err$(h) \leqslant \epsilon$.*

Démonstration.

$$\text{err}(h) = \mathbb{P}_{x \tilde{} \mathcal{D}}(x \models \phi_c \wedge x \not\models \phi_h) \leqslant \sum_{l \in \phi_h} \mathbb{P}_{x \tilde{} \mathcal{D}}(x \models \phi_c \wedge x \not\models l) \leqslant \sum_{l \in \phi_h} \frac{\epsilon}{2n} \leqslant \epsilon$$

□

Calculons ensuite la probabilité suivante :

$$\mathbb{P}_{x \tilde{} \mathcal{D}}(\text{Tous les mauvais littéraux apparaissent dans } S)$$

Soit l un mauvais littéral. On obtient alors :

$$\mathbb{P}_{x \tilde{} \mathcal{D}}(\forall x \in S, x \models \phi_c \implies x \models l) \leqslant (1 - \frac{\epsilon}{2n})^m \leqslant e^{-\frac{\epsilon m}{2n}} \text{ car } 1 + x \leqslant e^x$$

$$\mathbb{P}_{x \tilde{} \mathcal{D}}(\text{Il existe un mauvais littéral dans } S) \leqslant \sum_l \mathbb{P}(l \text{ apparaisse dans } S) \leqslant \sum_l e^{-\frac{\epsilon m}{2n}} \leqslant 2n.e^{-\frac{\epsilon m}{2n}}$$

En appliquant la définition de l'apprentissage (définition 4), on obtient l'inégalité suivante :

$$\delta \geqslant 2n.e^{-\frac{\epsilon m}{2n}}$$

Cette inégalité est vraie si $\boxed{m \geqslant \frac{2n}{\epsilon}(\ln(2n) + \ln(\frac{1}{\delta}))}$.

Conclusion
On a l'erreur qui est bien un polynôme en $\frac{1}{\epsilon}$. Cependant, ce n'est pas efficacement apprenable car m est en $\ln \frac{1}{\delta}$.