Learnability can be Undecidable

Travail encadré de recherche

Adrien Ragot
Professeurs encadrants : Myriam Quatrini, Pierre Pudlo

24 juin 2019

Aix Marseille Université

Contexte et Motivation

Apprenabilité EMX et compressibilité

Cardinalité et compression

Une classe de concept dont l’apprenabilité EMX n’est pas décidable, et conséquence sur les dimensions

Conclusion

Contexte et Motivation

Motivations

L'objet de ce travail a été de comprendre l'article "Learnability can be undecidable", [1].

L'article porte sur le domaine de l'apprentissage automatique. Lors de l'apprentissage automatique on cherche à apprendre des concepts, des éléments d'une classe $\mathcal{F}$ de parties du domaine X. Le point de départ est l'apprentissage PAC.

Definition (Apprenabilité PAC, Appreneur PAC)

Un appreneur PAC [1] de $\mathcal{F}$ est un algorithme $L : \bigcup_{d\in\mathbb{N}} L(X^d) \to \mathcal{F}$ tel que, pour toute distribution $\mathcal{D}$ sur X,

$$\forall c \in \mathcal{F}, \forall \varepsilon, \forall \delta, \exists d, \Pr_{S\sim\mathcal{D}^d}\left[\Pr_{x\sim\mathcal{D}}(L(S)(x) \neq h(x)) \leq \varepsilon\right] \geq 1-\delta$$

En fait il s'avère que l'on peut caractériser l'apprenabilité PAC d'une famille $\mathcal{F}$, par la compressibilité de la famille, mais aussi par sa dimension VC.

Alors la motivation de l'article est la question suivante ; peut on généraliser les caractérisations fondamentales de l'apprentissage PAC à d'autres types d'apprentissage ?

1. pour Probably Approximately Correct

L'apprentissage EMX

En particulier, les auteurs traitent de la question pour l'apprentissage EMX[2].

En fonction du type d'apprentissage on attend des conditions différentes sur l'hypothèse retournée, et la nature de l'appreneur peut varier (fonction ou algorithme).

Definition (Appreneur EMX)

On dit qu'une application $G : \bigcup_{\ell \in \mathbb{N}} X^\ell \to \mathcal{F}$ est un (ϵ, δ)-EMX appreneur, si et seulement si, Il existe un entier d tel que, pour toute mesure de probabilité P à support fini sur X, quel que soit $S = (S_1, \ldots, S_d)$ i.i.d.[3] une suite de d variables aléatoires suivant la loi P on a ;

$$\mathbb{P}_S \left[\underbrace{\mathrm{Opt}_P(\mathcal{F}) - \mathbb{E}_P(G(S))}_{\text{erreur}} \geq \epsilon \right] \leq \delta$$

avec $\mathrm{Opt}_P(\mathcal{F}) = \sup_{h \in \mathcal{F}} \mathbb{E}_P(h)$.

2. Pour Estimating the Maximum
3. indépendantes et identiquement distribuées

Apprenabilité EMX et compressibilité

Compression

L'objet de la première partie de l'article c'est donc de montrer l'équivalence entre apprenabilité EMX et compressibilité.

Remarque.
Là où la LW[4]-compressibilité caractérise l'apprenabilité PAC. C'est une autre notion de compressibilité qui caractérise l'apprenabilité EMX.

Definition (Schème de compression monotone)

On dit qu'une application $\eta : X^d \to \mathcal{F}$ est un $m \to d$ schème de compression monotone si ; Quelque soit S un échantillon de taille au plus m d'un $h \in \mathcal{F}$, il existe un sous échantillon $S' \subseteq S$ de taille d, tel que ; $S \subseteq \eta(S')$.

Si il existe un $m \to d$ schème de compression monotone de $\mathcal{F}$, on dit que $\mathcal{F}$ est $m \to d$ compressible.

Theorem

Si $\mathcal{F}$ est une famille dirigée d'ensembles finis.

1. $\mathcal{F}$ est EMX apprenable
2. Il existe un entier m, tel que $\mathcal{F}$ est $m + 1 \to m$ compressible.

4. pour Littlestone-Warmuth

Boosting

On montre d'abord ce qui est appelé le boosting, c'est à dire comment la compression faible entraîne la compression forte. En fait on montre préalablement ;

Theorem

$\mathcal{F}$ est m $\to$ d compressible $\Rightarrow$ $\mathcal{F}$ est m $+ 1 \to$ d compressible

Ce qui nous permet d'avoir l'équivalence suivante ;

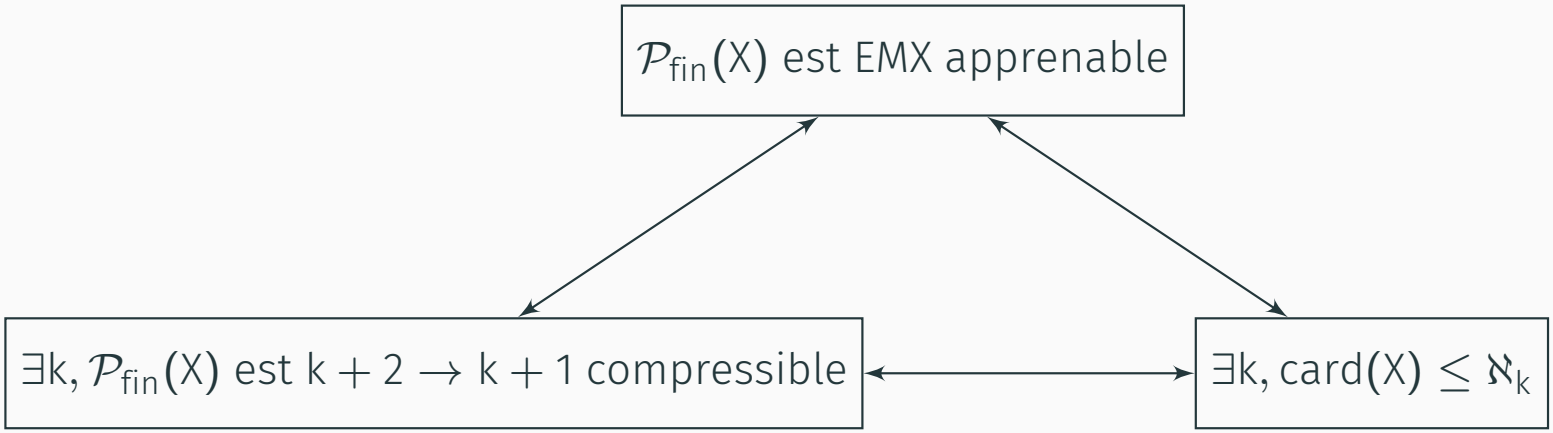
Theorem

Quel que soit d un entier, les assertions sont équivalentes

1. (Compressibilité faible) Il existe un entier m $>$ d tel que, $\mathcal{F}$ est m $\to$ d compressible
2. (Compréssibilité forte) Pour tout entier m, $\mathcal{F}$ est m $\to$ d compressible

Notons que pour montrer ces assertions on fera appelle à l'axiome du choix.

En fait lorsque le domaine X est de cardinal infini et qu'on s'interesse à l'apprenabilité de $\mathcal{P}_{\text{fin}}(X)$ on a même les 3 équivalences suivantes ;



Où $\aleph_k$ désigne le k-ième cardinal infini non dénombrable.

La compressibilité caractérise l'apprenabilité EMX

D'abord on montre les implications suivantes ;

Theorem

1. $\mathcal{F}$ est faiblement compressible $\Longrightarrow$ $\mathcal{F}$ est fortement compressible.
2. $\mathcal{F}$ est fortement compressible $\Longrightarrow$ $\mathcal{F}$ est EMX apprenable.
3. $\mathcal{F}$ est faiblement apprenable EMX $\Longrightarrow$ $\mathcal{F}$ est faiblement compressible.

Ce qui nous permettra de conclure l'équivalence entre l'apprenabilité EMX d'une famille $\mathcal{F}$ dirigée d'ensembles finis, et la compressibilité de cette famille.

On va s'attacher à expliciter la première implication, c'est à dire le boosting.

C'est à dire, comment à partir de η un scheme de compression $m \to d$ on en construit λ un scheme $m + 1 \to d$. On considère S un échantillon à $m + 1$ éléments, et donc la preuve consiste à ;

1. trouver S' un sous échantillon de S de taille d.
2. Puis une méthode λ telle que, $S \subseteq \lambda(S')$.

Boosting (idée de la preuve)

1. Considérons x un point de S. Alors, $S = T \cup \{x\}$. Et T est un échantillon de taille m, donc compressible par η en T′, c'est à dire, $T \subseteq \eta(T')$.

Puisque $d < m$, $T' \cup \{x\}$ est de taille au plus m, on peut donc compléter cette échantillon en un échantillon $R \subseteq S$ de taille m. Et R est alors compressible par η en S′ ; $R \subseteq \eta(S')$.

2. Alors on **choisi** ce que retourne λ comme un concept H qui contient tout les $\eta(V)$, où V est un sous échantillon de $\eta(S')$. Montrons que $H = \lambda(S')$ contient bien l'échantillon S.

- R est compressé par η en S′, donc $S' \subseteq R \subseteq \eta(S')$; alors il en suit $\eta(S') \subseteq \lambda(S')$, et donc de fait, $R \subseteq \lambda(S')$.
- Par ailleurs R complète $T' \cup \{x\}$ et donc, $T' \subseteq R \subseteq \eta(S')$, et donc il en suit que, $T \subseteq \eta(T') \subseteq \lambda(S')$.

Alors remarquons que T est un sous échantillon de $m + 1$ points de S ne contenant pas x, mais R lui contient x, alors on conclut ; $S \subseteq \lambda(S')$. □

Cardinalité et compression

Cardinalité et compression

Ici on veut expliciter le lien entre le cardinal de X et la compressibilité de $\mathcal{P}_{\text{fin}}(X)$.

Theorem

Quel que soit k un entier ;

$\mathcal{P}_{\text{fin}}(X)$ est $k+2 \to k+1$ compressible $\Leftrightarrow$ X est cardinal au plus $\aleph_k$.

On va s'attacher à expliciter le sens $(2) \Rightarrow (1)$de l'équivalence. Supposons X de cardinal $\aleph_k$, et montrons que $\mathcal{P}_{\text{fin}}(X)$ est $k+2 \to k+1$ compressible. On considère donc, S une partie de $k+2$ de points de X et on souhaite la compresser en un sous échantillon S' de taille $k+1$.

Remarque. si on explicite une méthode $f : X^{k+1} \to \mathcal{P}_{\text{fin}}(X)$, qui à S' retourne le singleton de $S \setminus S'$, alors on aura en fait trouvé un scheme de compression monotone. Puisqu'alors il suffit de poser $\eta : T \mapsto f(T) \cup T$ pour obtenir un scheme de compression. Cette méthode n'est valide que pour la famille $\mathcal{P}_{\text{fin}}(X)$, puisqu'elle réside sur le fait que si T est finie, T est un concept.

Pour construire f on va faire appel aux bons ordres. Par un corollaire de Zermelo, on peut assurer X est bien ordonnable par un ordre $<$, ceci tel que $(X, <)$ est isomorphe à son cardinal ici, $(\aleph_k, \in)$.

Itération

Alors on va se servir de la proposition suivante ; tout segment initial propre d'un cardinal α, est nécessairement de cardinal plus petit strictement que α.

Alors considérons $\max(S)$, puis $I_k = \{y \in X \mid y < \max(S)\{$. Remarquons deux faits ; $I_{\max(S)}$ contient les $k+1$ points restants de S ie, $S_k = S \setminus \{\max(S)\}$. Mais aussi, I_k est de cardinal strictement plus petit que $\aleph_k$.

On peut réitérer alors exactement la même "opération" sur l'échantillon restant S_k, alors on obtient un segment initial I_{k-1} qui contient les k points restant S_{k-1} et est de cardinal strictement plus petit que $\aleph_{k-1}$.

On fait l'opération k fois on obtient alors, S_0 qui contient 1 élément, et est de cardinal strictement plus petit que $\aleph_0$ le dénombrable. C'est à dire que S_0 est fini donc dans la classe des concepts $\mathcal{P}_{\text{fin}}(X)$.

Alors remarquons que à partir des $k+1$ points de $S \setminus S_0$ on peut construire S_0 par l'itération précédente. Donc en fait on a trouvé une f telle que $S_0 \subseteq f(S \setminus S_0)$. Ceci quel que soit un échantillon S de taille $k+2$. Donc par la remarque on en déduit l'existence d'un $k+2 \to k+1$ scheme de compression monotone de $\mathcal{P}_{\text{fin}}(X)$.

Une classe de concept dont l’apprenabilité EMX n’est pas décidable, et conséquence sur les dimensions

Apprenabilité EMX non décidable

On a lié apprenabilité, compression et cardinalité. Certains énoncés sur les cardinaux sont connus pour être non décidables donc par les équivalences on va certainement pouvoir obtenir des énoncés non décidables, et en effet si on fixe l'ensemble domaine comme $X = [0, 1]$ on va obtenir un énoncé où la puissance du continue apparait.

$$
\begin{aligned}
\mathcal{P}_{\text{fin}}([0,1]) \text{ est EMX apprenable} \quad &\Leftrightarrow \quad \text{Il existe un entier k tel que,} \mathcal{P}_{\text{fin}}([0,1]) \\
&\qquad \text{est } k+2 \to k+1 \text{ compressible.}^{5} \\
&\Leftrightarrow \quad \text{Il existe un entier k tel que,} \\
&\qquad [0,1] \text{ est de cardinal au plus } \aleph_k. \\
&\Leftrightarrow \quad \text{Il existe un entier k tel que, } 2^{\aleph_0} \leq \aleph_k \\
&\Leftrightarrow \quad 2^{\aleph_0} < \aleph_\omega.^{6}
\end{aligned}
$$

Definition. Une formule ϕ est dites indécidable relativement à une théorie T, si il existe $\mathfrak{M}_0$ et $\mathfrak{M}_1$ deux modèles de T telle que, $\mathfrak{M}_0 \vDash \phi$ et $\mathfrak{M}_1 \vDash \neg\phi$.

Et la formule $2^{\aleph_0} < \aleph_\omega$ n'est pas décidable dans ZFC. De fait puisque deux formules équivalentes ont la même valeur de vérité dans un même modèle, l'apprenabilité de $\mathcal{P}_{\text{fin}}(X)$ n'est pas décidable.

5. le cas d'un scheme $1 \to 0$ est exclu pour un ensemble de cardinalité infini.
6. ω désigne le premier cardinal infini, c'est à dire le dénombrable.

Indécidabilité et dimension

On a donc vu que comme l'apprentissage PAC, l'apprentissage EMX peut être caractérisée par une compression. Reste la question de la dimension. Premièrement, définissons donc ce qu'on entends par dimension.

1. (Caractérise l'apprenabilité) $\mathcal{F}$ est EMX-apprenable $\Leftrightarrow \mathcal{D}(\mathcal{F})$ est finie.
2. (Invariant par Modèle) Si l'assertion "$\mathcal{D}(\mathcal{F})$ est finie" est vrai dans un modèle de ZFC, elle est vrai dans tout modèle de ZFC.

Alors puisque deux formules équivalentes ont la même valeur de vérité dans un même modèle il est clair qu'il en suit que l'assertion "$\mathcal{F}$ est EMX-apprenable" est aussi invariante par modèle, mais ceci pour tout famille $\mathcal{F}$ de concept.

Mais pourtant on a explicité pour la famille $\mathcal{F} = \mathcal{P}_{\text{fin}}(X)$ que son apprenabilité n'était pas la même dans tous les modèles. Ca veut donc aussi dire que $\mathcal{D}(\mathcal{P}_{\text{fin}}(X))$ est de dimension finie dans $\mathfrak{M}_0$, et $\mathcal{D}(\mathcal{P}_{\text{fin}}(X))$ est de dimension infinie dans $\mathfrak{M}_1$. Mais alors on vient contredire notre notion de dimension.

C'est pour cela qu'il n'y en fait pas de notion de dimension pour l'apprentissage EMX.

Conclusion

Pourquoi les apprentissage EMX et PAC diffèrent dans leurs caractérisations ?

Les différences remarquables entres les deux types d'apprentissage qui pourraient être susceptibles de générer les différences entre les deux apprentissages ;

1. l'étiquetage ; les échantillons de l'apprentissage EMX et PAC ne sont pas les mêmes, les échantillons de l'apprentissage PAC sont des suites de points indexés par un concept, c'est à dire qu'on indique si le point est dans un concept ou non (par exemple dans le cas des rectangles, on a l'information qui indique si le point est dans le rectangle ou pas). Là où les échantillons pour l'EMX sont simplement des suites de points d'un concept.
2. Nature de l'appreneur ; dans le cas EMX c'est une fonction, tandis que pour l'apprentissage PAC l'appreneur est un algorithme.

Conclusion

L'article à ainsi montrer que les caractérisations de l'apprentissage PAC ne sont pas généralisables aux autres types d'apprentissage de façon certaines, ceci en traitant le cas de l'apprentissage EMX.

L'intêret de l'apprentissage EMX c'est que ses appreneurs sont des fonctions, on a donc pas à considérer l'aspect calculatoire qu'aurait un algorithme appreneur de l'apprentissage PAC. Et aussi de fait le traitement de l'apprentissage EMX est plus ensembliste, mais comme on a pu le voir ceci à un coup ; notamment l'indécidabilité de l'apprenabilité EMX.


Shai Ben-David , Pavel Hrubeš , Shay Moran , Amir Shpilka and Amir Yehudayoff Learnability can be undecidable Link Article, Nature Machine Intelligence, 2019.


Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff On a learning problem that is independent of the set theory ZFC axioms Link Article, 2019.


Nick Littlestone, Manfred K. Warmuth Relating Data Compression and Learnability Link Technical Report, University of California, 1986.