Machine Learning 4: PAC learning, No-Free-Lunch theorem, uniform convergence

Master 2 Computer Science

Aurélien Garivier

2018-2019

UNIVERSITÉ DE LYON | ENS DE LYON

Table of contents

IA Challenge 2019


Accueil FAQ Se connecter S'inscrire Défi 2018

Airbus Defence and Space et les établissements toulousains (INSA , Université Paul Sabatier , Toulouse School of Economy) s'associent pour proposer un défi en reconnaissance d'images à des étudiants de Master 2 orientés "mathématiques pour l'IA".

Construit sur le principe d'un concours Kaggle , l'objectif de ce défi est d'identifier au mieux la présence ou l'absence d'une éolienne sur une image satellite à l'aide d'un algorithme d'apprentissage automatique.

Dans une démarche d'innovation ouverte et de soutien à l'entreprenariat, Airbus Defence and Space propose l' accès à de très nombreux types et d'exemples d' images satellite afin de susciter l'innovation et la création de nouveaux services comme lors d' hackathons ou de Challenges.

Dans le présent défi, identifier la présence d'une éolienne est un premier pas pour, par exemple, évaluer rapidement et automatiquement l'importance des parcs éoliens et donc de cette ressource énergétique à travers le monde, ainsi que son évolution. Pour atteindre cet objectif, Airbus Defence and Space met à disposition un ensemble de 85 000 échantillons d'images SPOT avec la présence, ou non, d'une éolienne; 68 000 sont disponibles pour l'apprentissage, les autres 17 000 sont utilisées en test pour le classement public et celui final.

Registration password: MVog!RFB

Use ENSL email address and keep me informed.

PAC learning

PAC learnability: "probably approximately correct"

Definition

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_m$ such that for every $\epsilon, \delta \in (0,1)$, for every distribution D_X on $\mathcal{X}$ and for every labelling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, D_X, f$ then when $S = \big((X_1, f(X_1)), \ldots, (X_m, f(X_m))\big)$ with $(X_i)_{1 \leq i \leq m} \overset{iid}{\sim} D_X$,

$$\mathbb{P}\Big(L_{(D_X,f)}(\hat{h}_m) \geq \epsilon\Big) \leq 1 - \delta$$

for all $m \geq m_{\mathcal{H}}(\epsilon, \delta)$.
The smallest possible function $m_{\mathcal{H}}$ is called the *sample complexity* of learning $\mathcal{H}$.

Remark: Valiant's PAC requires also sample complexity and running time polynomial in $1/\epsilon$ and $1/\delta$.

Finite hypothese classes are PAC-learnable

The sample complexity of finite hypothese classes in the realizable case is smaller than $m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$:

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Let $\epsilon, \delta \in (0, 1)$ and let m be an integer that satisfies

$$m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon} .$$

Then, for any labeling function f and for any distribution D_X on $\mathcal{X}$, under the realizability assumption, with probability at least $1 - \delta$ over the choice of iid sample S of size m, any ERM hypothesis $\hat{h}_m$ is such that

$$L_{(D_X, f)}(\hat{h}_m) \leq \epsilon .$$

Proof

The realizability assumption implies that an ERM $\hat{h}_S$ has empirical risk $L_S(\hat{h}_S) = 0$. Hence,

$$
\begin{aligned}
\mathbb{P}\left(L(\hat{h}_S) \geq \epsilon\right) &= D_X^{\otimes m}\left(\left\{S \in \mathcal{X}^m : \exists h \in \mathcal{H}, L_S(h) = 0 \text{ and } L_D(h) \geq \epsilon\right\}\right) \\
&= D_X^{\otimes m}\left(\bigcup_{h : L_D(h) \geq \epsilon} S_h\right) \quad \text{where } S_h = \left\{S \in \mathcal{X}^m : L_s(h) = 0\right\} \\
&\leq \sum_{h : L_D(h) \geq \epsilon} D_X^{\otimes m}(S_h) \\
&= \sum_{h : L_D(h) \geq \epsilon} \prod_{i=1}^{m} \underbrace{D_X(\{x \in \mathcal{X} : h(x) = f(x)\})}_{= 1 - L_D(h) \leq 1 - \epsilon} \\
&\leq \sum_{h : L_{(D_X, f)}(h) \geq \epsilon} \prod_{i=1}^{m} (1 - \epsilon) \leq |\mathcal{H}|(1 - \epsilon)^m \leq |\mathcal{H}| \exp(-m\epsilon)\ .
\end{aligned}
$$

This quantity is smaller than δ for $m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$.

Agnostic PAC learnability

Definition

A hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $S \mapsto \hat{h}_m$ such that for every $\epsilon, \delta \in (0,1)$, for every distribution D on $\mathcal{X} \times \mathcal{Y}$ when $S = \big((X_1, Y_1), \dots, (X_m, Y_m)\big) \overset{iid}{\sim} D$,

$$\mathbb{P}\Big(L_D(\hat{h}_m) \geq \min_{h' \in \mathcal{H}} L_D(h') + \epsilon\Big) \leq 1 - \delta$$

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

The smallest possible function $m_{\mathcal{H}}$ is called the *sample complexity* of learning $\mathcal{H}$.

If the realizable assumption holds, boils down to PAC learnability.

Otherwise, recall that the best Bayes classifier reaches $\min_{h' \in \mathcal{H}} L_D(h')$.

Learning via uniform convergence

Definition

A training set S is called ϵ-representative (wrt domain $\mathcal{X} \times \mathcal{Y}$, hypothese class $\mathcal{H}$, loss function l and distribution D) if

$$\forall h \in \mathcal{H}, \left| L_S(h) - L_D(h) \right| \leq \epsilon \, .$$

Lemma

If S is $\epsilon/2$-representative, then any ERM $\hat{h}_m$ defined by $\hat{h}_m \in \arg\min_{h \in \mathcal{H}} L_S(h)$ satisfies:

$$L_D\left(\hat{h}_m\right) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon \, .$$

Proof: for every $h \in \mathcal{H}$,

$$L_D\left(\hat{h}_m\right) \leq L_S\left(\hat{h}_m\right) + \frac{\epsilon}{2} \leq L_S\left(h\right) + \frac{\epsilon}{2} \leq L_D\left(h\right) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \, .$$

Uniform Convergence Property

Definition

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* (wrt $\mathcal{X} \times \mathcal{Y}$ and l) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every distribution D over $\mathcal{X} \times \mathcal{Y}$, a sample $S = \big((X_1, Y_1), \ldots, (X_m, Y_m)\big) \overset{iid}{\sim} D$ of size $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ has probability at least $1 - \delta$ to be ϵ-representative.

Corollary

If $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$, then $\mathcal{H}$ is agnostically PAC learnable with a sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$. Furthermore, the ERM is a successful PAC learner for $\mathcal{H}$.

Finite classes are agnostically PAC-learnable

Theorem

Let $\mathcal{H}$ be a finite hypothesis class. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

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

Moreover, $\mathcal{H}$ is agnostically PAC learnable using an ERM algorithm with sample complexity

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

Proof: Hoeffding's inequality and the union bound.

No-Free-Lunch theorems: when learning is not possible

The No-Free-Lunch theorem

Theorem

Let A be any learning algorithm for binary classification over a domain $\mathcal{X}$. If the training set size is $m \leq |\mathcal{X}|/2$, then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

- there exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_D(f) = 0$;
- with probability at least $1/7$ over the choice of $S \sim \mathcal{D}^{\otimes m}$,

$$L_{\mathcal{D}}\big(A(S)\big) \geq \frac{1}{8} \,.$$

Note that the ERM over $\mathcal{H} = \{f\}$,or over any set $\mathcal{H}$ such that $m \geq 8 \log(7|\mathcal{H}|/6)$, is a successful learner in that setting.

Proof

Take $C \subset \mathcal{X}$ of cardinality $2m$, and $\{0,1\}^C = \{f_1, \ldots, f_T\}$ where $T = 2^{2m}$. For each $1 \leq i \leq T$, we denote by D_i the probability distribution on $C \times \{0,1\}$ defined by

$$D_i(\{x,y\}) = \begin{cases} \frac{1}{2m} \text{ if } y = f_i(x) , \\ 0 \text{ otherwise.} \end{cases}$$

We will show that $\max_{1\leq i\leq T} \mathbb{E}\left[L_{D_i}(A(S))\right] \geq 1/4$, which entails the result thanks to the small lemma: if $P(0 \leq Z \leq 1) = 1$ and $\mathbb{E}[Z] = 1/4$, then $\mathbb{P}(Z \geq 1/8) \geq 1/7$. Indeed, $1/4 \leq \mathbb{E}[Z] \leq \mathbb{P}(Z < 1/8)/8 + \mathbb{P}(Z \geq 8) = 1/8 - 7\mathbb{P}(Z \geq 8)/8$.

All the X-samples $S_1, \ldots, S_k$, for $k = m^{2m}$ are equaly likely. For $1 \leq j \leq k$, if $S_j = (x_1, \ldots, x_m)$ we denote by $S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m))$.

$$\begin{aligned} \max_{1\leq i\leq T} \mathbb{E}\left[L_{D_i}(A(S))\right] &= \max_{1\leq i\leq T} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \\ &= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \min_{1\leq j\leq k}\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \ . \end{aligned}$$

Fix $1 \leq j \leq k$, denote $S_j = (x_1, \ldots, x_m)$ and define $\{v_1, \ldots, v_p\} = C \setminus \{x_1, \ldots, x_m\}$, where $p \geq m$. Then

$$L_{D_i}(A(S_j^i)) = \frac{1}{2m}\sum_{x\in C} \mathbb{1}\{A(S)(x) \neq f_i(x)\} \geq \frac{1}{2p}\sum_{r=1}^{p} \mathbb{1}\{A(S)(v_r) \neq f_i(v_r)\}$$

and hence

$$\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}\mathbb{1}\{A(S)(v_r) \neq f_i(v_r)\} \geq \frac{1}{2}\min_{1\leq r\leq p}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}\{A(S)(v_r) \neq f_i(v_r)\} \ .$$

Fix $1 \leq r \leq p$. Then the functions $\{f_i : 1 \leq i \leq T\}$ can be grouped into $T/2$ pairs of functions $(\tilde{f}_i^0, \tilde{f}_i^1)$, $1 \leq i \leq T/2$ which agree on all $x \in C$ except on v_r, and for all $1 \leq i \leq T/2$ it holds that $\mathbb{1}\{A(S)(v_r) \neq \tilde{f}_i^0(v_r)\} + \mathbb{1}\{A(S)(v_r) \neq \tilde{f}_i^1(v_r)\} = 1$.

Hence, $\sum_{i=1}^{T} \mathbb{1}\{A(S)(v_r) \neq f_i(v_r)\} = \sum_{i=1}^{T/2} \mathbb{1}\{A(S)(v_r) \neq \tilde{f}_i^0(v_r)\} + \mathbb{1}\{A(S)(v_r) \neq \tilde{f}_i^1(v_r)\} = T/2$, which concludes the proof.

Consequence: Curse of Dimensionality

Theorem

Let $c > 1$ be a Lipschitz constant. Let A be any learning algorithm for binary classification over a domain $\mathcal{X} = [0,1]^d$. If the training set size is $m \leq (c+1)^d/2$, then there exists a distribution $\mathcal{D}$ over $[0,1]^d \times \{0,1\}$ such that:

- $\eta(x)$ is c-Lipschitz;
- the Bayes error of the distribution is 0;
- with probability at least $1/7$ over the choice of $S \sim \mathcal{D}^{\otimes m}$,

$$L_{\mathcal{D}}\big(A(S)\big) \geq \frac{1}{8}\ .$$

Uniform convergence for infinite classes: VC dimension

Shattering

Definition

Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and let $C = \{c_1, \dots, c_m\} \subset \mathcal{X}$. The *restriction* of $\mathcal{H}$ to C is the set of functions $C \to \{0,1\}$ that can be derived from $\mathcal{H}$:

$$\mathcal{H}_C = \Big\{ (c_1, \dots, c_m) \to \big(h(c_1), \dots, h(c_m)\big) : h \in \mathcal{H} \Big\} .$$

Shattering

A hypothesis class $\mathcal{H}$ *shatters* a finite set $C \subset \mathcal{X}$ if $\mathcal{H}_C = \{0,1\}^C$.

Example:

- $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$.
- $\mathcal{H}^2_{\mathrm{rec}} = \big\{ h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1 \text{ and } a_2 \leq b_2 \big\}$ where

$$h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2\,; \\ 0 & \text{otherwise}\,. \end{cases}$$

VC dimension

Definition

The *Vapnik Chervonenkis dimension* $\mathrm{VCdim}(\mathcal{H})$ of a hypothesis class $\mathcal{H}$ is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathrm{VCdim}(\mathcal{H}) = \infty$.

Theorem

Let $\mathcal{H}$ be a class of infinite VC-dimension. Then $\mathcal{H}$ is not PAC-learnable.

Proof: for every training size m, there exists a set C of size $2m$ that is shattered by $\mathcal{H}$. By the NFL theorem, for every learning algorithm A there exists a probability distribution D over $\mathcal{X} \times \{0, 1\}$ such that $L_D(h) = 0$ but with probability at least $1/7$ over the training set, we have $L_D\big(A(S)\big) \geq 1/8$.