De Laplace à Vapnik : les fondements mathématiques de l'apprentissage statistique

Nicolas Vayatis

21 janvier 2026

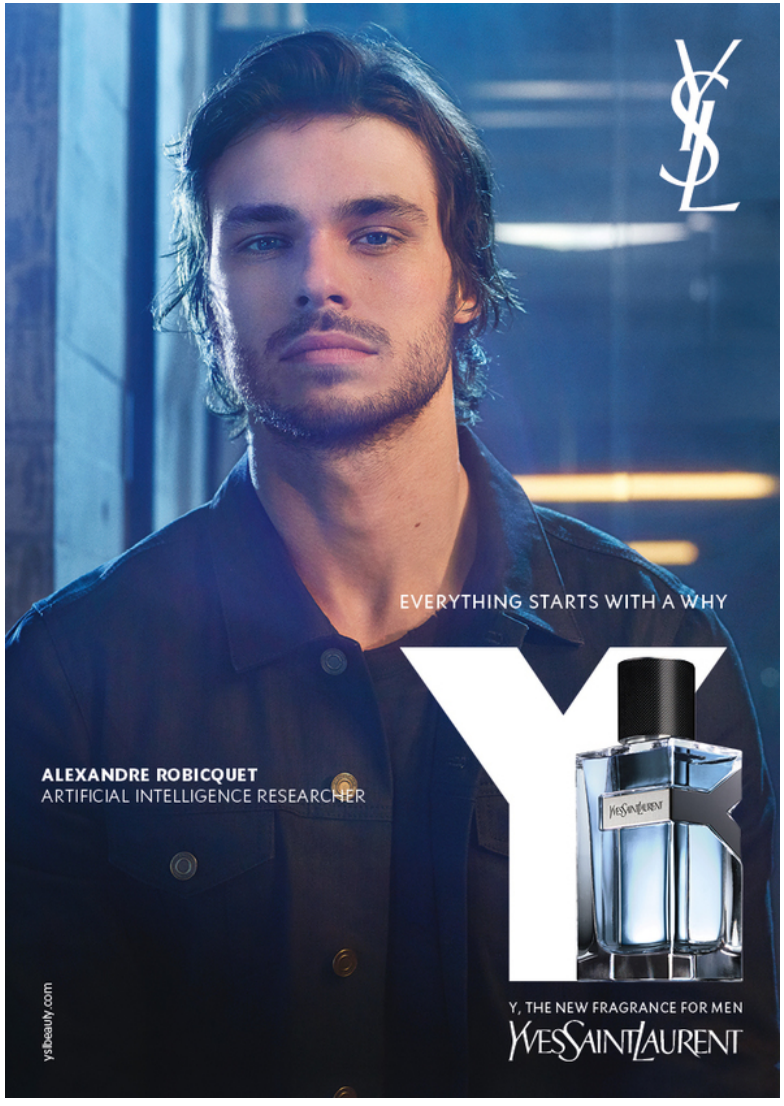
Maths en Herbe, IHES

Que faire de son temps ?

Travailler et jouer !

Faire des maths, bon sang !

Mais oui ! Les maths mènent à tout !



Et un mathématicien, ça fait quoi ?

Exemples de défis intellectuels, esthétiques...

- "Travailler jusqu'à l'épuisement total, mais pas plus !" (M. Talagrand)
- Démontrer des théorèmes
- Construire des preuves (très) techniques
- Rechercher le beau (une "belle" preuve, un "joli" problème)
- Raconter une histoire
- Servir les autres (mathématiques, sciences, technologies, société)
- ...

Et cet exposé ? ? ?

Les personnages principaux

Pierre-Simon de Laplace (1749-1827)

Vladimir Vapnik (1936-...)

Du côté de Laplace

La méthode de Laplace (1/2)

- Soit un compact $K \subset \mathbb{R}^d$ et deux fonctions $g, f : K \to \mathbb{R}$ telles que, $\forall t \in \mathbb{R}$,

$$\mathcal{I}(t) = \int_K f(x) e^{-tg(x)} dx \quad \text{est bien définie.}$$

- L'approximation (basique) de Laplace fournit un équivalent de $\mathcal{I}(t)$ quand $t \to +\infty$

$$\mathcal{I}(t) \sim \frac{f(x^*)}{\sqrt{\det g''(x^*)}} \left(\frac{2\pi}{t}\right)^{d/2} e^{-tg(x^*)}$$

sous les hypothèses suivantes :

- f est continue et g est C^2 sur K
- g a un minimum global strict x^* dans l'intérieur de K
- $f(x^*) \neq 0$.

La méthode de Laplace (2/2)

- Schéma de preuve :
  - On peut supposer : $g(x^*) = 0$, $g''(x^*) = \mathrm{Id}$ et $x^* = 0$
  - Développement de Taylor à l'ordre 2 de g
  - Changement de variable :

$$h(x) = (g(x)/\|x\|^2)\mathbb{I}\{x \neq 0\} + (1/2)\mathbb{I}\{x = 0\}$$

  - Le résultat découle du théorème de convergence dominée et de la formule :

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}u^T u\right)\, du = (2\pi)^{d/2}$$

- Exemple : formule de Stirling

$$n! = \Gamma(n+1) = \int_0^\infty x^n e^{-x} dx \sim n^{n+1}\sqrt{\frac{2\pi}{n}}e^{-n} = \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

En probabilités : fonction génératrice des moments

- Soit Z une variable aléatoire réelle de densité f absolument continue par rapport à la mesure de Lebesgue, la fonction génératrice des moments est :

$$M(t) = \mathbb{E}(e^{tZ}) = \int_{\mathbb{R}} e^{tz} f(z)\, dz$$

  pour tout t tel que l'intégrale existe.

- Propriétés de base :
  - La fonction M caractérise la loi de Z.
  - Si 0 est dans l'intérieur du domaine de définition de M, alors : $M^{(k)}(0) = \mathbb{E}(Z^k)$ (moment d'ordre k)

Bornes célèbres sur la fonction génératrice des moments

- Cas de variables aléatoires bornées (régime sous-gaussien) : soit $\mathbb{P}(Z \in [a, b]) = 1$, $\mathbb{E}(Z) = 0$ alors

$$\mathbb{E}(e^{tZ}) \leq \exp\left(\frac{t^2(b-a)^2}{8}\right), \ \forall t > 0$$

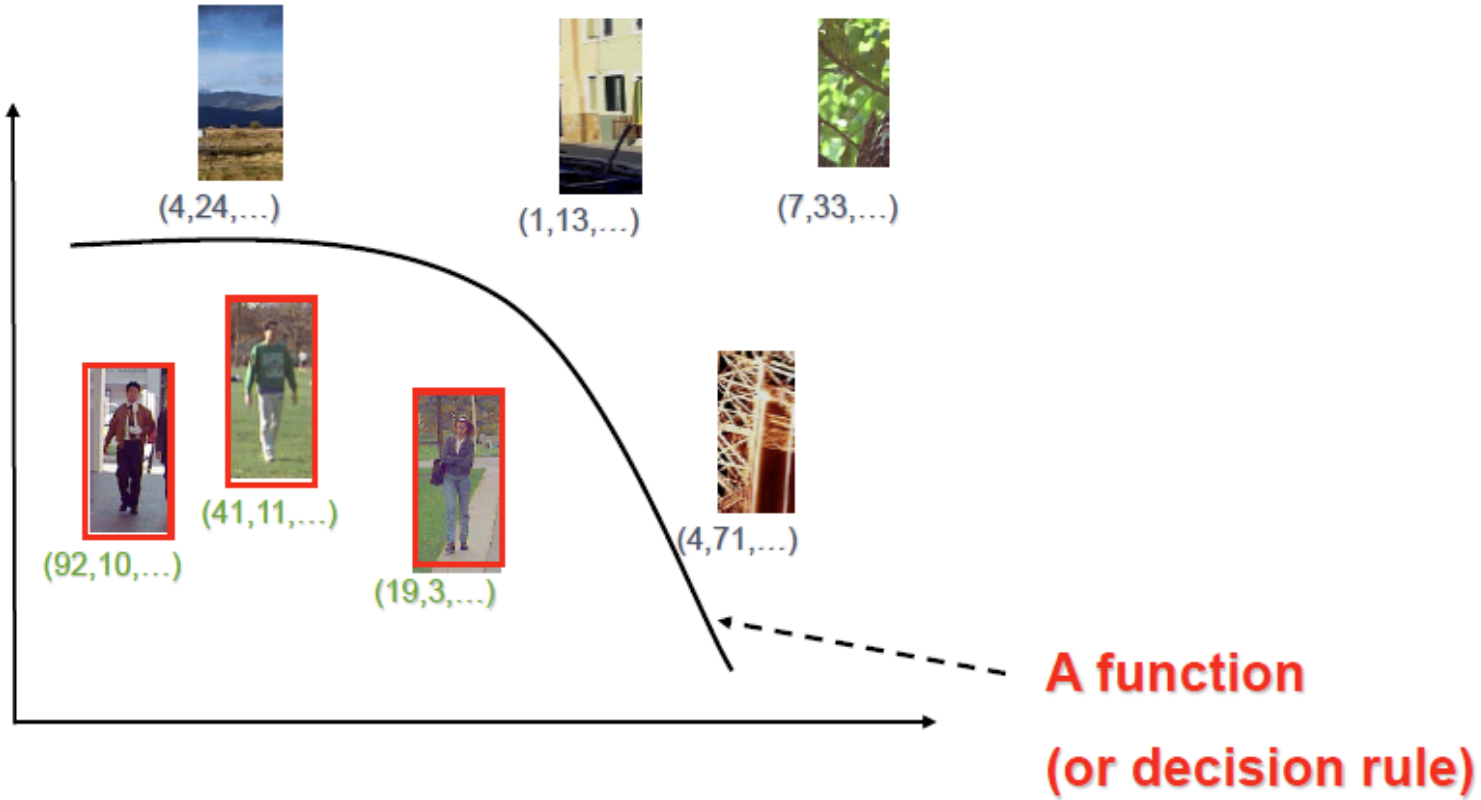
- Cas de variables aléatoires bornées avec variance explicite (régime poissonien) : soit $\mathbb{P}(|Z| \leq c) = 1$, $\mathbb{E}(Z) = 0$ et $\mathbb{E}(Z^2) = \sigma^2$ alors

$$\mathbb{E}(e^{tZ}) \leq \exp\left(\frac{\sigma^2}{c^2}(e^{tc} - 1 - tc)\right), \ \forall t > 0$$

De Laplace aux voitures sans conducteur

Détection de piéton dans les images

Détection de piéton dans les images



Formalisation du problème

- Pour démontrer la capacité de généralisation de la règle de décision f prise dans une classe $\mathcal{F}$, on doit assurer la propriété suivante :

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}_P\big(f(Z)\big) \right| \xrightarrow{P} 0$$

où $Z, Z_1, \ldots, Z_n$ sont IID de loi P et $\mathbb{E}_P(f(Z)) = \int f(z) dP(z)$

- De manière plus générale, on se pose la question de la convergence uniforme de la loi empirique vers la *vraie* loi :

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \xrightarrow{P} 0$$

où $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}$ et δ_Z mesure de Dirac

Loi des grands nombres

- Convergence en probabilité (définition) - Une suite de variables aléatoires $(U_n)_{n\geq 1}$ converge en probabilité vers une variable aléatoire U si :

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}(|U_n - U| > \varepsilon) = 0$$

- Loi des grands nombres (version faible) - Supposons f bornée et $Z, Z_1, \ldots, Z_n$ IID tels que $\mathbb{E}_P(Z^2) < \infty$, alors :

$$\frac{1}{n}\sum_{i=1}^{n} f(Z_i) \xrightarrow{P} \mathbb{E}_P\big(f(Z)\big)$$

- Borne de la réunion : soit A et B deux événements, alors

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Inégalité de Hoeffding

- Consider $U, U_1, \ldots, U_n$ IID over $[0,1]$. Then, for any $\varepsilon > 0$:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}(U) > \varepsilon\right) \leq \exp(-2n\varepsilon^2)$$

- Proof argument : Chernoff's bounding method

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}(U) > t\right)$$
$$\leq \exp\left(-\sup_{s>0}\left(nst - n\log\mathbb{E}\left(e^{s(U-\mathbb{E}(U))}\right)\right)\right)$$

Loi uniforme des grands nombres

- Si $\mathcal{F}$ est de cardinal fini :
  on a (borne-de-la-réunion + inégalité de Hoeffding) : $\forall \varepsilon > 0$

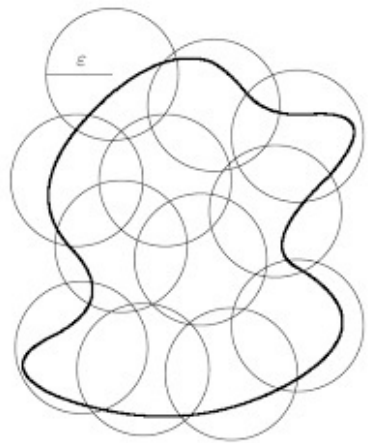
$$\mathbb{P}\left(\max_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}_P\big(f(Z)\big) \right| > \varepsilon \right)$$

$$\leq |\mathcal{F}| \max_{f \in \mathcal{F}} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}_P\big(f(Z)\big) \right| > \varepsilon \right) \leq 2|\mathcal{F}| \exp(-2n\varepsilon^2)$$

- Si $\mathcal{F}$ est dénombrable : l'argument s'effondre.
- Question : comment quantifier la taille d'une classe fonctionnelle ?

Selon Kolmogorov...

- Soit une classe de fonctions $\mathcal{F}$ muni d'une métrique $\|\cdot\|$
- Un ε-recouvrement $\mathcal{T}$ est un ensemble d'éléments de $\mathcal{F}$ tels que $\forall f \in \mathcal{F}$ il existe un élément $h \in \mathcal{T}$ tel que $\|h - f\| \leq \varepsilon$



- Le nombre de couverture $N(\varepsilon)$ est le cardinal du plus petit ε-recouvrement de $\mathcal{F}$

Lois uniformes des grands nombres sous contrôle métrique

- Si les fonctions de $\mathcal{F}$ sont uniformément bornées par M, on a :

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}|P_n(f)-P(f)|>\varepsilon\right)\leq N\left(\frac{\varepsilon}{8M}\right)\exp\left(-\frac{n\varepsilon^2}{2M^2}\right)$$

[D. Pollard (1984)]

- Si $N(\varepsilon)\sim\varepsilon^{-\kappa}$ avec $\kappa>0$ (e.g. $\kappa=d$ si $\mathcal{F}$ est paramétré par compact dans $\mathbb{R}^d$), alors on a :

$$\sup_{f\in\mathcal{F}}|P_n(f)-P(f)|\xrightarrow{P}0$$

- Mais hypothèses (trop) abstraites et contraignantes, vitesses peu explicites...

Du côté de Vapnik

Dénombrement dans un ensemble infini

- Soit $S \subset \mathbb{R}^d$ avec $|S| < +\infty$ et $\mathcal{C}$ une famille de sous-ensembles de $\mathbb{R}^d$.
- On dit que $\mathcal{C}$ pulvérise S si pour tout $S' \subset S$, il existe $C \in \mathcal{C}$ tel que $S' = S \cap C$.
- La dimension de Vapnik-Chervonenkis (VC) de $\mathcal{C}$ est définie par

$$V := \sup\{|S| : S \text{ est pulvérisé par } \mathcal{C}\}$$

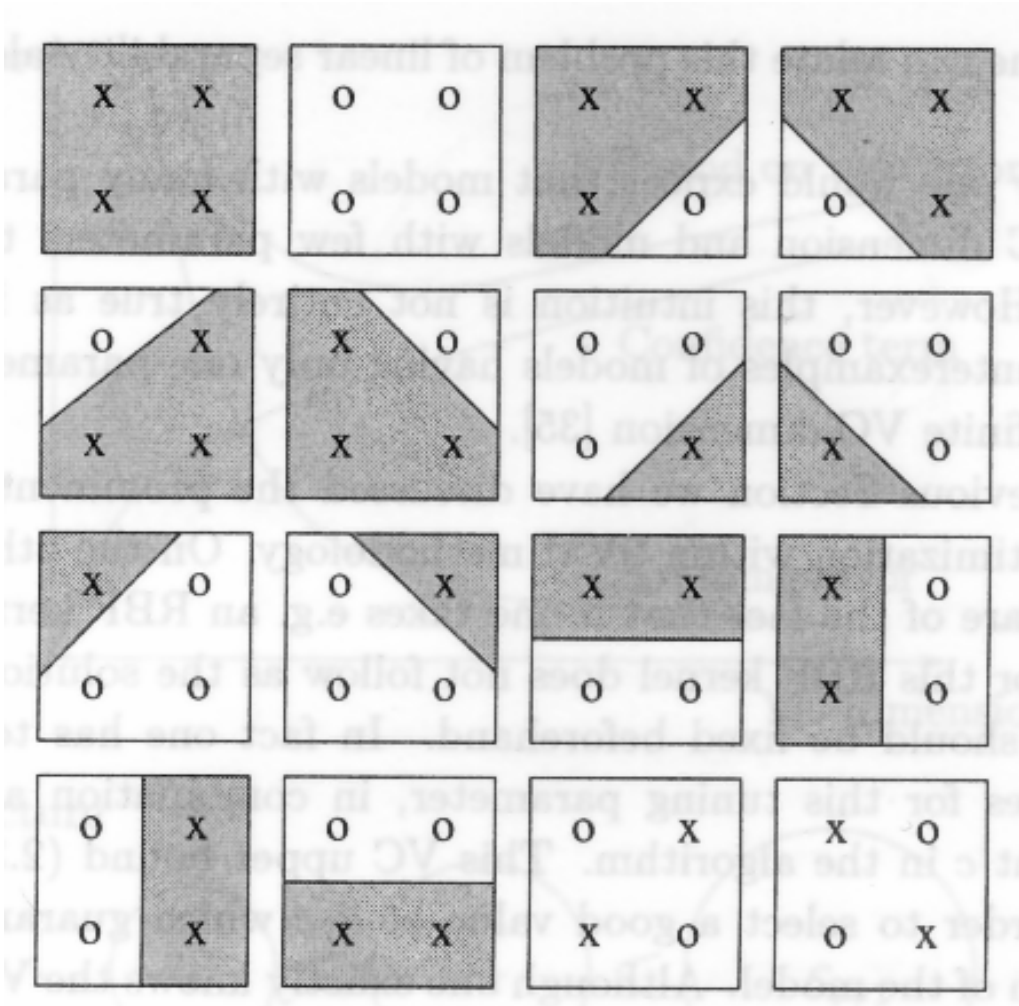
- Par conséquent, pour tout S, tel que $|S| \leq V$, on a :

$$|\{S \cap C : C \in \mathcal{C}\}| = 2^{|S|}$$

- Lemme de Sauer-Shelah : soit $\mathcal{C}$ de VC dimension V

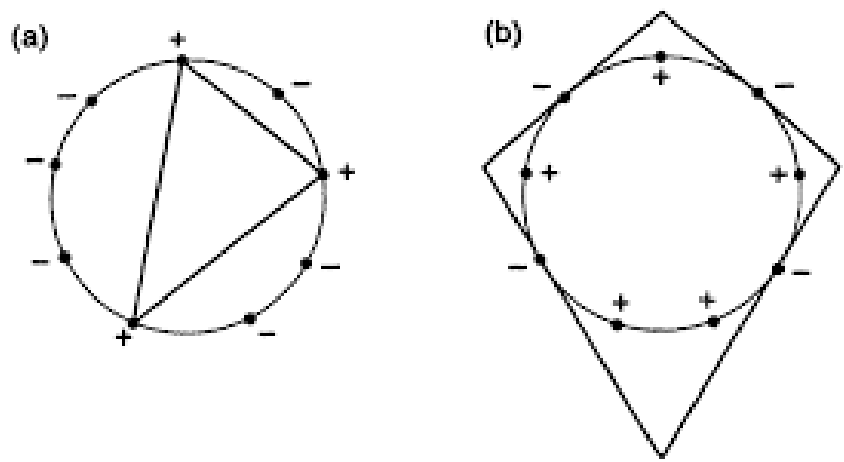
$$s(\mathcal{C}, n) = \max_{S \,:\, |S|=n} |\{S \cap C : C \in \mathcal{C}\}| \leq \sum_{k=0}^{V} \binom{n}{k} \leq (n+1)^V$$

VC dimension des demi-plans



Exemples de calculs exacts de VC dimension

- Demi-espaces dans $\mathbb{R}^d$: $V = d + 1$
- Rectangles de côtés parallèles aux axes dans $\mathbb{R}^2$: $V = 4$
- Tous les rectangles dans $\mathbb{R}^2$: $V = 7$
- Triangles dans $\mathbb{R}^2$: $V = 7$
- Polygones Convexes dans $\mathbb{R}^2$: $V = +\infty$



Le nombre de paramètres ne quantifie pas la complexité !

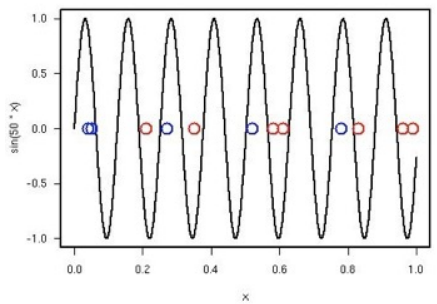
- Ensemble de fonctions indicatrices paramétrées par un seul paramètre :

$$h(x) = \mathbb{I}\{x : \sin(\omega x) > 0\} \ , \ \text{where } \omega \in [0, 2\pi)$$

- La VC dimension de l'ensemble est infinie ! En prenant l'ensemble de points $\{x_j = 2\pi 10^{-j} \ : \ j = 1, \ldots, n\}$ et en considérant toutes les labellisations binaires possibles $\{y_1, \ldots, y_n\}$, le choix du paramètre

$$\widehat{\omega}_n(y_1, \ldots, y_n) = \frac{1}{2}\left(1 + \sum_{i=1}^{n}\left(\frac{1 - y_i}{2}\right)10^i\right)$$

montre qu'on peut pulvériser cet ensemble.



Lois uniformes des grands nombres sous contrôle combinatoire

- Si les fonctions de $\mathcal{F}$ sont des fonctions indicatrices indexées par une classe d'ensembles $\mathcal{C}$ de VC dimension V, on a :

$$\mathbb{P}\left(\sup_{C\in\mathcal{C}} |P_n(C) - P(C)| > \varepsilon\right) \leq 8s(\mathcal{C}, n) \exp\left(-\frac{n\varepsilon^2}{128}\right)$$
$$\leq 8(n+1)^V \exp\left(-\frac{n\varepsilon^2}{128}\right)$$

[Vapnik, Chervonenkis (1968)]

- On en déduit avec probabilité au moins $1-\delta$

$$\sup_{C\in\mathcal{C}} |P_n(C) - P(C)| \leq \sqrt{\frac{128V\log(n+1)}{n}} + \sqrt{\frac{128\log(8/\delta)}{n}}$$

- Super ! mais peut mieux faire...

Vitesses dans les lois uniformes des grands nombres

En travaillant un peu mieux...

- Inégalité de concentration de Mc Diarmid (plus fort que Hoeffding !)
- Inégalité de Dudley (basée sur la technique du chaînage)

... on obtient, avec probabilité au moins $1-\delta$:

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(Z_i)-\mathbb{E}_P\big(f(Z)\big)\right|\leq 192\sqrt{\frac{(3+\log 2)V}{n}}+2\sqrt{2\frac{\log(2/\delta)}{n}}$$

- On peut alors fournir un intervalle de confiance sur la performance en moyenne future d'un algorithme d'apprentissage qui a appris à détecter des piétons sur un échantillon de taille finie dès lors que sa VC dimension est finie.
- On a fini ? Problème résolu ?

Commentaires

- **La question du biais.** Les garanties fournies portent sur les fluctuations dues au caractère aléatoire des données d'apprentissage (erreur d'estimation) mais quid de la capacité de représentation de la famille $\mathcal{C}$ de la meilleure fonction de décision (erreur d'approximation) ?

  Question ouverte : Réconcilier les bornes d'erreurs en estimation et approximation pour le problème de classification.

- **Calibration de la complexité.** Est-ce que la VC dimension est la notion de complexité pertinente ? Quels conseils aux praticiens ?

  Question ouverte : Comment calibrer a priori la complexité de l'apprentissage ? Intervalles de confiance numériquement plausibles ?

Une notion géométrique de la complexité

- Soit deux échantillons de variables aléatoires indépendants l'un de l'autre : $X_1, \ldots, X_n$ IID sur $\mathbb{R}^d$ et un échantillon $\varepsilon_1, \ldots, \varepsilon_n$ IID telles que $\mathbb{P}(\varepsilon_1 = 1) = \mathbb{P}(\varepsilon_1 = -1) = 1/2$.
- La complexité (empirique) de Rademacher d'une classe de functions $\mathcal{F}$:

$$\widehat{R}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \middle| X_1, \ldots, X_n \right)$$

- Bornes sur la complexité de Rademacher :
  - Classe linéaire à coefficients bornés par M : $O(M/\sqrt{n})$
  - Classe de VC dimension finie : $O(\sqrt{V/n})$
  - Enveloppe convexe d'une classe de VC dimension finie : $O(\sqrt{V/n})$! ! !

Où Laplace rencontre Vapnik

Retour à Laplace : des vitesses spécifiques

- Soit Z une variable aléatoire suivant une loi de Bernoulli de paramètre $p \in (0,1)$

$$\mathbb{E}(e^{tZ}) = pe^t + (1-p),\ \forall t > 0$$

- Méthode de Chernoff appliquée à la moyenne empirique :

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - p > \varepsilon\right) \leq \exp\left(-nH(p+\varepsilon, p)\right)$$

où $H(q,p) = q\log(q/p) + (1-q)\log((1-q)/(1-p))$

- Cette vitesse est asymptotiquement exacte en échelle logarithmique (Théorème de Crámer) et on a :

$$\forall \varepsilon > 0,\ \lim_{p\to 1} \frac{1}{\varepsilon^2} H(p+\varepsilon, p) = +\infty$$

Retour à Vapnik : approcher la complexité effective

- **Conjecture.** Soit $\mathcal{C}$ de VC dimension finie et $\mathcal{P}$ classe de mesures sur $\mathcal{B}(\mathbb{R}^d)$

$$\sup_{P\in\mathcal{P}} \mathbb{P}\left(\sup_{C\in\mathcal{C}} |P_n(C) - P(C)| > \varepsilon\right)$$
$$\simeq K(n\varepsilon^2)^{v-1/2} \exp\left(-nH(p^* + \varepsilon, p^*) \wedge H(p^* - \varepsilon, p^*)\right)$$

où $p^* = \underset{q=P(C)\ :\ C\in\mathcal{C}, P\in\mathcal{P}}{\arg\min} |q - 1/2|$

- Idée : valider empiriquement cette conjecture (avant de la démontrer...)
- Challenges numériques même dans un cas trivial (2D, loi uniforme, demi-espaces) :
  - Simulation d'événements rares par échantillonnage préférentiel
  - Calcul du supremum par dualité projective

Calculs de VC dimension effective

Digit	degree	V_{th}	V_{eff}
0	3	$\sim 10^6$	530
1	7	$\sim 10^{16}$	101
2	3	$\sim 10^6$	842
3	3	$\sim 10^6$	1157
4	4	$\sim 10^9$	962
5	3	$\sim 10^6$	1090
6	4	$\sim 10^9$	626
7	5	$\sim 10^{12}$	530
8	4	$\sim 10^9$	1145
9	5	$\sim 10^{12}$	1226

Vapnik, 1998

q	V	δV	K	$\delta K/K$
1.0	1.00	0.22	0.95	23%
0.9	0.99	0.20	0.97	23%
0.8	0.99	0.14	0.95	15%
0.7	1.03	0.14	0.90	16%
0.6	1.06	0.18	0.80	20%
0.5	0.98	0.14	0.67	16%
0.4	0.78	0.09	0.53	10%
0.3	0.68	0.06	0.42	7%
0.2	0.62	0.04	0.33	5%
0.1	0.57	0.04	0.23	4%

Vayatis, 2000

Voir aussi : Vapnik-Levin-LeCun, 1994

Conclusions et messages personnels

- La réponse à la conjecture est peut-être du côté de la méthode de Laplace
- Exemple de réflexions mathématiques à partir de problèmes concrets
- Utilisation de la simulation numérique pour tenter d'invalider des conjectures
- Toutes les sciences seront mathématiques ou ne seront pas !
- Remerciements chaleureux aux personnages principaux de *mon* histoire : Robert, Pascal, Gábor, Sacha, Jean-Michel

Pour en savoir plus : quelques lectures saines

- P. Barbe / Approximation of integrals over asymptotic sets with applications to statistics and probability / ArXiv, 2003
- K. W. Breitung / Asymptotic approximations for probability integrals / Springer, 1994
- R. van Handel / Probability in high dimension / Princeton preprint, 2016
- R. Vershynin / High dimensional probability / Cambridge University Press, 2018
- S. Boucheron, G. Lugosi, P. Massart / Concentration Inequalities : A Nonasymptotic Theory of Independence / Oxford University Press, 2016
- V. Vapnik / Statistical Learning Theory / Wiley, 1998
- F. Bach / Learning Theory from First Principles / MIT Press, 2024

Backup

Plus fort que Hoeffding ! la concentration

[McDiarmid's inequality] Consider $Z_1, \ldots, Z_n$ IID. Under a regularity assumption on the function f called the *bounded difference assumption* with constant c/n, we have, for any $t > 0$

$$\mathbb{P}\left(f(Z_1, \ldots, Z_n) - \mathbb{E}\big(f(Z_1, \ldots, Z_n)\big) > t\right) \leq \exp\left(-2nt^2/c^2\right)$$

and

$$\mathbb{P}\left(f(Z_1, \ldots, Z_n) - \mathbb{E}\big(f(Z_1, \ldots, Z_n)\big) < -t\right) \leq \exp\left(-2nt^2/c^2\right)$$

- Here the average of IID random variables is replaced by a general function of these IID variables.
- Take-home message : **Independence is more important/general than averaging**

Bounded difference assumption

- Consider a function f of n variables. We say that f has bounded differences if the variations along each variables are uniformly bounded.
- Here we need to have : for some $c > 0$

$$\sup_{z_1,\ldots,z_n,z_i'} |f(z_1,\ldots,z_n) - f(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n)| \leq \frac{c}{n}$$