

R

ENDEZ-VOUS

P.58 Logique & calcul
P.64 Idées de physique
P.68 Chroniques de l'évolution
P.72 Science & gastronomie
P.74 À picorer

DESSINER UN ÉLÉPHANT AVEC UN SEUL PARAMÈTRE

Reproduire n'importe quel jeu de données en nombre fini, par exemple des points qui esquissent la silhouette d'un animal: c'est l'exploit que permet une fonction paramétrisée par un unique nombre réel.

L'AUTEUR



JEAN-PAUL DELAHAYE
professeur émérite
à l'université de Lille
et chercheur au
laboratoire Cristal
(Centre de recherche
en informatique, signal
et automatique de Lille)



Jean-Paul Delahaye
a notamment publié:
**Les Mathématiciens
se plient au jeu,**
une sélection de ses
chroniques parues
dans *Pour la Science*
(Belin, 2017).

Supposons que l'on dispose de données et que l'on recherche une explication simple à ces données, ou au moins un modèle qui permette de les reproduire. On partira par exemple de valeurs de mesures qui, pour $x = 0, 1, \dots, n$, donnent $y_0, y_1, \dots, y_n$. Si l'on trouve une fonction f qui vérifie $f(0) = y_0, f(1) = y_1, \dots, f(n) = y_n$, on sera assez satisfait. On utilisera f pour prédire $f(n+1), f(n+2), \dots$ et l'on analysera f pour comprendre les données étudiées et peut-être pour formuler une théorie les expliquant.

Bien évidemment, on recherche des fonctions f les plus simples possible. Les polynômes, c'est-à-dire les fonctions de la forme $f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$ sont parmi les plus simples. On sera donc assez satisfait si l'on trouve un polynôme qui reproduit $y_0, y_1, \dots, y_n$.

Or le monde mathématique est généreux et permet souvent de faire tout ce que l'on veut. Un joli résultat, dénommé «théorème d'interpolation de Lagrange», indique que quelles que soient les données $y_0, y_1, \dots, y_n$, il existe un polynôme de degré au plus n qui redonne $y_0, y_1, \dots, y_n$ pour $x = 0, 1, \dots, n$ (voir l'encadré 2). Mais ce polynôme est un surajustement (ou *overfitting*): on se donne trop de paramètres, ici les coefficients du polynôme, et ce que l'on trouve n'a pas vraiment d'intérêt car on était certain d'y arriver!

Aussi le calcul du polynôme d'interpolation de Lagrange ne permettra que très rarement d'anticiper ce qui se passe pour $n+1, n+2, \dots$, et ce polynôme ne sera pas utile sur un plan

théorique. Heureusement, il se peut que le polynôme trouvé soit de faible degré. Si c'est le cas, cela signifie que sa découverte aurait résulté de la prise en compte de moins de données (autant que le degré du polynôme, plus une unité) et donc que le polynôme retrouve les autres données. Dans une telle situation, il ne sera pas absurde de croire que le polynôme calculé permet d'anticiper $f(n+1), f(n+2), \dots$.

Le risque de surajustement est un problème important dans le domaine de l'apprentissage profond en intelligence artificielle, car il conduit à des réseaux de neurones qui sont bien adaptés à un grand ensemble de données... et à rien d'autre.

À propos du choix des paramètres dans un modèle, le mathématicien d'origine hongroise John von Neumann a fait une remarque amusante: «Avec quatre paramètres, je peux faire un éléphant, et avec cinq, je peux faire remuer sa trompe.»

UN SEUL PARAMÈTRE SUFFIT POUR TOUT AJUSTEMENT

Un article récent de Steven Piantadosi, du Département de psychologie de l'université de Californie à Berkeley, a montré que von Neumann aurait pu remplacer le nombre «quatre» de sa phrase par le nombre «un». L'article est bien sûr un jeu, mais ce qu'il montre, en mettant en œuvre des mathématiques assez simples, est remarquable et doit servir de mise en garde: un seul paramètre permet d'ajuster une fonction pour qu'elle redonne mille données ou un milliard si l'on s'y prend bien.

UN SEUL PARAMÈTRE POUR L'ÉLÉPHANT

Steven Piantadosi a proposé la fonction : $f_p(x) = \sin^2(2^{kx} \arcsin(\sqrt{p}))$, où le nombre réel p est un paramètre, et a montré qu'elle présente une extraordinaire propriété d'ajustement. La valeur de k est fixée une fois pour toutes, et indique la précision avec laquelle on approchera les données. On considère par exemple des valeurs $y_0, y_1, \dots, y_n$ telles que les points de coordonnées $(0, y_0), (1, y_1), \dots, (n, y_n)$ dessinent un éléphant, en serrant assez les abscisses et en dessinant des points assez gros.

La méthode de Steven Piantadosi (expliquée dans

les encadrés 3 et 4) permet alors de déterminer la valeur du paramètre p pour que : $f_p(0) = y_0, f_p(1) = y_1, \dots, f_p(n) = y_n$, autrement dit pour que le graphe de $f_p(x)$ dessine un éléphant. Un seul paramètre suffit pour avoir un éléphant ! Bien sûr, on peut obtenir tous les dessins qu'on veut en changeant la valeur de p . Laurent Boué a écrit des programmes disponibles sur internet qui permettent le calcul de p pour les dessins souhaités et a fait les calculs pour un oiseau et une tortue (voir la bibliographie).

L'énoncé démontré par Steven Piantadosi est le suivant.

- On se fixe une précision ϵ , par exemple $1/1000$, ce qui signifie que l'on recherchera à avoir des égalités avec une erreur inférieure à ϵ .
- Il existe une fonction $f_p(x)$ dépendant d'un paramètre réel p , telle que, quelles que soient les données $y_0, y_1, \dots, y_n$ en nombre fini, il existe une valeur p du paramètre pour laquelle $f_p(0) = y_0, f_p(1) = y_1, \dots, f_p(n) = y_n$, avec une erreur inférieure à ϵ pour chaque égalité.

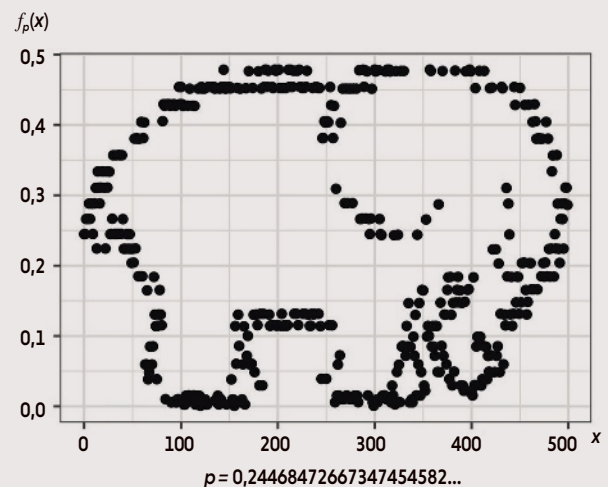
En clair, en faisant varier le paramètre p , la fonction $f_p(x)$ s'ajuste avec une précision ϵ à tout jeu de données fixé à l'avance, aussi grand soit-il.

La fonction miraculeuse de Steven Piantadosi est : $f_p(x) = \sin^2(2^{kx} \arcsin(\sqrt{p}))$. Rappelons que $\arcsin(x)$ est la fonction inverse de $\sin(x)$, ce qui signifie que pour tout x compris entre $-\pi/2$ et $\pi/2, y = \sin(x) \Leftrightarrow x = \arcsin(y)$.

La valeur k qui apparaît dépend de la précision ϵ souhaitée, mais une fois ϵ fixée, k l'est aussi. Bien évidemment, le chercheur a calculé le paramètre p qui donne un éléphant. L'encadré 1 montre le résultat.

D'OÙ VIENT L'EXPLOIT ?

Comprendre en détail comment la fonction $f_p(x)$ réalise cet exploit, et comment s'y prendre pour trouver les bonnes valeurs du paramètre p , est un jeu mathématique qui exige un peu de calcul (voir les encadrés 3 et 4). Cependant, l'idée générale est simple : on va cacher dans le paramètre p qui, parce qu'il est un nombre réel, peut contenir une infinité d'informations, les nombres $y_0, y_1, \dots, y_n$. Ensuite, la



$p = 0,74933466...$



$p = 0,70704013...$