Une brève histoire de l'apprentissage

Gilbert Saporta

Chapitre 1

Une brève histoire de l'apprentissage

Gilbert Saporta

1.1 Introduction

La statistique est souvent présentée comme « *the science of learning from data* » (Franklin et Agresti [2013]). En intelligence artificielle le terme d'apprentissage ou *Machine Learning* renvoie à des techniques algorithmiques fondées sur l'adaptation de modèles à des données dans un but essentiellement prédictif. Malgré des vocabulaires parfois différents et des communautés scientifiques au départ distinctes, la convergence de ces deux disciplines définit ce que l'on appelle l'apprentissage statistique. On lira avec intérêt le best seller de Hastie *et al.* [2009].

Une manière de donner à des machines de véritables capacités d'apprentissage (et non seulement de reproduction comme les automates de Vaucanson) est de modéliser l'apprentissage humain. Représenter l'homme comme une machine permet alors de concevoir des machines apprenantes. On peut faire remonter les prémices de l'apprentissage au 17$^{\text{ème}}$ siècle avec la théorie de l'animal-machine de René Descartes (1596-1650), étendue par La Mettrie à l'homme-machine (1748). L'article History of Artificial Intelligence de Wikipedia remonte jusqu'à l'antiquité avec les statues de dieux parlantes et cite les machines (théoriques !) du philosophe catalan Ramon Lull (1235-1315) conçues comme des systèmes mécaniques produisant des résultats logiques. Gottfried Leibniz (1646-1716) fut influencé par les travaux de Lull. On notera avec intérêt que dans un de ses derniers entretiens publié après son décès, McCulloch [1974] indiquait « *I conclude that cybernetics really starts with Descartes rather than with Leibniz* ». Laissons cependant les débats sur la préhistoire de l'apprentissage aux historiens des sciences et consacrons-nous à l'époque moderne.

Après les débuts du connexionnisme, contemporain de l'invention de l'ordinateur, on fait souvent référence à la période 1956-1974 comme l'âge d'or

de l'apprentissage machine. Il fut suivi de ce qu'il est convenu d'appeler le « premier hiver » d'une dizaine d'années, causé par le pessimisme sur les performances des réseaux de neurones (voir plus loin) et l'arrêt de nombreux programmes de recherche. La décennie des années 80 vit le développement des systèmes experts et le renouveau du connexionnisme avec la redécouverte de l'algorithme de rétropropagation, suivi du « deuxième hiver » de 1987 à 1993, qui prit fin avec l'apparition des SVM et des réseaux récurrents. Depuis les années 2000, on assiste avec le *deep learning* et ses succès technologiques à un regain d'intérêt pour le connexionnisme. Cette progression pendulaire, selon l'expression de Philippe Besse, a concerné surtout le monde de l'informatique, tandis que du côté de la statistique, les évolutions furent plus douces. Sans doute est-ce dû au moindre rôle des subventions.

1.2 Des neurones artificiels au perceptron

1.2.1 Le premier modèle et les cybernéticiens

L'article fondateur est sans conteste celui de McCulloch et Pitts [1943] intitulé « *A Logical Calculus of Ideas Immanent in Nervous Activity* » qui eut une influence décisive en proposant le premier modèle mathématique d'un neurone artificiel (associable en réseaux) avec une fonction de réponse à seuil puisque, selon les auteurs, l'activité nerveuse fonctionne par tout ou rien. L'article, outre un traitement mathématique rigoureux, établit des liens avec la calculabilité au sens de la machine de Turing et se termine par des considérations sur la causalité. La figure **??** du chapitre **??** du présent ouvrage est une représentation moderne du neurone de McCulloch et Pitts.

Warren Sturgis McCulloch (1898-1969) était un neurophysiologiste américain et Walter Harry Pitts, Jr. de 25 ans son cadet (1923-1969) un logicien spécialisé en neurosciences. Leurs travaux les rapprochèrent du fondateur de la cybernétique Norbert Wiener (1894-1964), qui écrivit que Pitts était « *without question the strongest young scientist whom I have ever met... I should be extremely astonished if he does not prove to be one of the two or three most important scientists of his generation, not merely in America but in the world at large* ».

Norbert Wiener fut un enfant prodige qui obtint son doctorat à 17 ans à Harvard. Sa famille prétendait descendre du philosophe et savant juif Maïmonide (1138-1204).

L'ouvrage essentiel de la cybernétique est « *Cybernetics or Control and Communication in the Animal and the Machine* » que Wiener publia en 1948, simultanément chez Hermann à Paris et aux presses du MIT (Wiener [1948]). Le mot cybernétique issu du grec *kubernetes* était déjà utilisé par Platon dans ses dialogues ; il provient du verbe grec *kubernân* (piloter un char ou un bateau) d'où dérivent gouvernail, gouvernance, etc. La fortune du mot cybernétique est

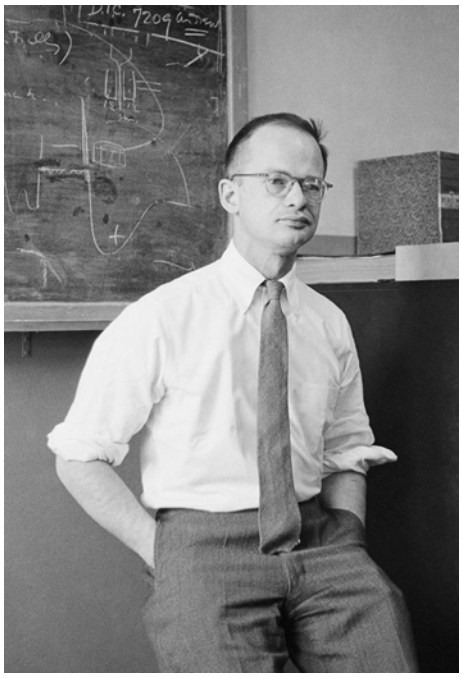

Figure 1.1 : *Walter Harry Pitts.*

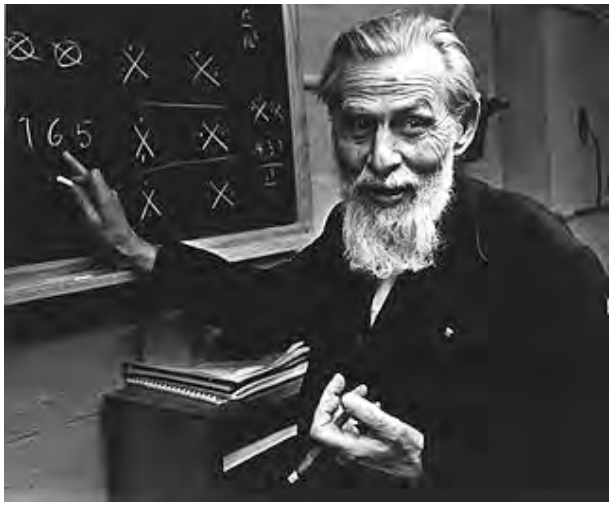

Figure 1.2 : *Warren Sturgis McCulloch.*

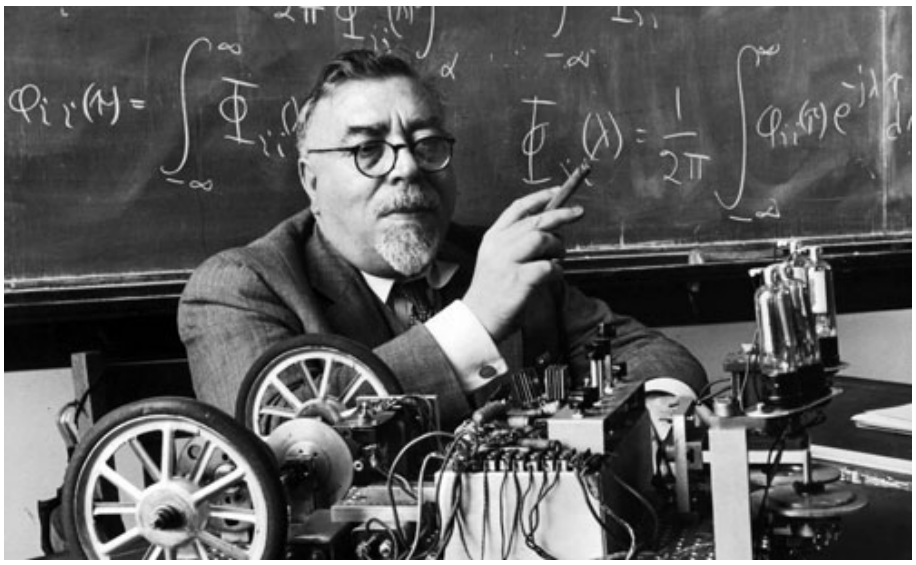

Figure 1.3 : *Norbert Wiener.*

intéressante : devenu quelque peu obsolète en français (il n'en reste guère que le préfixe cyber, et l'Association française pour la cybernétique économique et technique (AFCET) a disparu en 1998), il subsiste encore de nombreux départements et facultés de cybernétique dans les pays d'Europe centrale et orientale, en particulier en Russie. Après avoir méprisé cette science bourgeoise du temps de Staline et Lyssenko, l'Union Soviétique en fit une priorité à partir de la fin des années 50 et Norbert Wiener en devint un héros.

Fin 1943 Norbert Wiener emmena Pitts à une conférence qu'il organisait à Princeton avec John von Neumann (1903-1957), qui fut également impressionné par les capacités de Pitts. Ainsi se forma, autour de Wiener, Pitts, McCulloch, Lettvin et von Neumann, le groupe connu sous le nom des cybernéticiens.

En 1947 à la Second Cybernetic Conference, Pitts annonça qu'il écrirait sa thèse sur les réseaux de neurones probabilistes à trois dimensions. Mais en 1952 Norbert Wiener mit fin brutalement à la collaboration avec McCulloch et Pitts par un télégramme adressé à Jerry Wiesner, associate director of MIT's Research Laboratory of Electronics « *Please inform [Pitts and Lettvin] that all connection between me and your projects is permanently abolished. They are your problem. Wiener.* » Il n'adressa plus jamais la parole à Pitts et ne donna aucune explication. Pitts en fut profondément affecté, brula tous ses papiers de thèse, et sombra dans l'alcool : il décéda dans la solitude, des conséquences d'une cirrhose du foie en mai 1969, précédant de peu McCulloch. La rupture entre Wiener et McCulloch et Pitts resta longtemps un mystère. Dans leur ou-

Figure 1.4 : *John von Neumann.*

vrage « *Dark Hero of the Information Age: In Search of Norbert Wiener, the Father of Cybernetics* » (Conway *et al.* [2005]) publié en 2005 et traduit en français en 2012 (Conway *et al.* [2012]), Flo Conway et Jim Siegelman affirment que cette rupture fut le résultat d'un complot ourdi par l'épouse de Norbert Wiener, Margaret Engemann. Le mariage en 1926 avec Margaret Engemann, jeune immigrante allemande, avait été arrangé par les parents de Norbert Wiener. Margaret Wiener avait des ambitions mondaines et souhaitait tenir salon comme épouse d'un grand savant, mais elle ne supportait pas le style de vie bohême de McCulloch. Elle aurait déclaré à son époux qu'un et même plusieurs des jeunes chercheurs réunis autour de McCulloch lors de soirées arrosées avaient fait des avances à leur fille ainée, quatre ans auparavant.

1.2.2 Le connexionnisme et la naissance des ordinateurs

Les débuts de l'informatique sont inséparables du connexionnisme.

Quand en juin 1945, John von Neumann décrivit le principe de l'ordinateur programmable dans son historique « *First Draft of a Report on the EDVAC* », son unique référence bibliographique fut l'article de McCulloch et Pitts (Von Neumann [1945]). Pour réaliser cette machine, von Neumann suggérait d'utiliser des tubes à vide et de les relier selon le schéma des réseaux de McCulloch et Pitts. L'EDVAC (*Electronic Discrete Variable Automatic Computer*) fut mis en service en 1951. Contrairement à l'ENIAC (*Electronic Numerical Integrator Analyser and Computer*) qui opérait en décimal, l'EDVAC était un ordinateur binaire. Avec près de 6 000 tubes à vides et 12 000 diodes, il pesait près de 8 tonnes et nécessitait trois équipes de trente personnes qui se succèdaient en continu pour le faire fonctionner. Sa mémoire représentait à peine 5,5 ko.

Un rapport très peu connu intitulé « *Intelligent machinery* » prouve que

Figure 1.5 : *Mark I.*

sous le nom de « *unorganised machinery* » Alan Turing (1912-1954) avait également imaginé en 1948 un ordinateur fondé sur un réseau de type neurones connectés. Ce rapport a été redécouvert en 1968 et est disponible à l'adresse `http://www.alanturing.net/turing_archive/archive/l/l32/l32.php`.
Il ne mentionne pas l'article de McCulloch et Pitts qui, lui, citait les travaux de Turing. Il semble pourtant clair d'après Copeland et Proudfoot [1996] que Turing ayant fréquenté Norbert Wiener et John von Neumann connaissait les travaux de McCulloch et Pitts, mais on ne peut savoir quelle fut leur influence.

1.2.3 Le perceptron

L'autre grand précurseur des réseaux de neurones fut Frank Rosenblatt (1928-1971) connu comme l'inventeur du perceptron, la première « machine à apprendre » qui était à la fois une machine bien réelle et un algorithme. Le Mark 1 perceptron fut créé en 1957-58 selon les principes des réseaux multicouches à réponse linéaire à seuil. Rappelons que les neurones de McCulloch et Pitts étaient à réponse binaire (fonction d'Heaviside). Mark 1 était un classifieur visuel avec une couche d'entrée de 400 cellules photoélectriques selon une grille 20x20 simulant une rétine, une couche intermédiaire de 512 unités, et une couche de sortie de 8 unités. On peut voir la machine au Smithsonian Institute à Washington (figure **??**).

Frank Rosenblatt introduisit le terme de rétropropagation pour décrire sa méthode : *back-propagating error correction*. Parmi ses nombreuses publications, on citera son ouvrage fondamental (Rosenblatt [1962]) : « *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* ». Le perceptron de Rosenblatt ne pouvait s'appliquer qu'à des situations séparables

Figure 1.6 : *Frank Rosenblatt et le capteur de Mark I.*

linéairement (avec une indétermination de la solution) et l'enthousiasme retomba vite. Le livre de Marvin Minsky (1927-2016), condisciple de Rosenblatt, et Seymour Papert (1928-2016) « *Perceptrons* » (Minsky et Papert [1969]) publié en 1969 qui relevait ces limitations, fut utilisé à leur corps défendant pour discréditer pendant 10 ans les recherches sur les réseaux de neurones conduisant au « premier hiver de l'intelligence artificielle ». On fit grand cas de l'impuissance du perceptron à représenter le « ou exclusif » ou XOR , ce qui n'est vrai que pour un neurone isolé, mais pas pour un réseau, ainsi que l'impossibilité de séparer les spirales imbriquées illustrant la couverture du livre de Minsky et Papert.

Dans sa version d'origine, l'algorithme du perceptron est en effet un classifieur linéaire qui ne peut bien entendu pas traiter les cas non-linéairement séparables. L'utilisation d'une fonction d'activation non linéaire dans un réseau multicouche a résolu ensuite la difficulté, mais également les SVM. Frank Rosenblatt décéda peu après la parution du livre de Minsky et Papert, le 11 juillet 1971, le jour même de ses 43 ans, d'un accident de bateau en solitaire dans la baie de Chesapeake ; certains affirmèrent sans preuve que c'était un suicide. Un prix Frank Rosenblatt est décerné tous les ans depuis 2006 par l'IEEE ; parmi les lauréats on note T. Kohonen en 2008, J. Hopfield en 2009, V. Vapnik en 2012 et G. Hinton en 2014.

La redécouverte de l'algorithme de rétropropagation en 1986 par Rumelhart, les travaux de Hopfield, les succès en reconnaissance de la parole et de caractères, suscitèrent un regain d'intérêt pour les réseaux de neurones jusqu'à ce que certains nomment le « deuxième hiver de l'intelligence artificielle » 1987-1993, mais cet « hiver » ne fut pas si froid que cela. Dans cette période de nombreux travaux sur les réseaux de neurones furent publiés, et le concept de réseaux

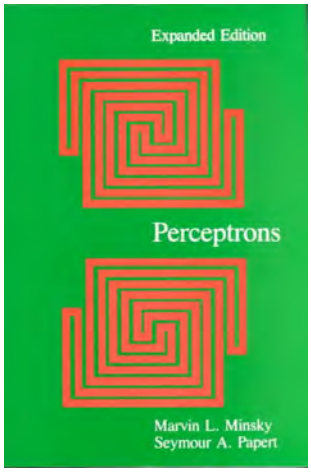

Figure 1.7 : *Minsky et Papert, Perceptrons, deuxième édition.*

Figure 1.8 : *Médaille en l'honneur de Frank Rosenblatt.*

convolutifs fut introduit. Une recherche active se développa en France autour de Françoise Fogelman-Soulié à l'Ecole des Hautes Etudes en Informatique de l'université Paris-Descartes, voir (Fogelman-Soulie *et al.* [1987] qui tissa des liens avec l'équipe de Larry Jackel aux Bell Labs. Par ailleurs, les approches par quantification vectorielle Kohonen [1989] ont connu de nombreux développements à cette période. On voit aussi les premières applications sérieuses en reconnaissance d'image : par exemple *Eigenfaces* utilisé par Turk et Pentland [1991] ou de parole Waibel *et al.* [1989] et Bottou [1991], ainsi que des résultats théoriques sur les capacités d'approximation des réseaux de neurones Cybenko [1989], la convergence de l'apprentissage, la régularisation.

L'intérêt pour les réseaux de neurones s'épuise plutôt vers 1992-1995, car d'une part il n'y a plus d'avancées théoriques, et d'autre part les capacités de traitement et de stockage sont à l'époque encore insuffisantes pour obtenir des résultats très convaincants, sauf sur des applications de niche (ex : lecture des codes postaux). Par ailleurs, il est (et reste) difficile de coder ces modèles (vitesse de calcul, tours de main pour accélérer la convergence, critères d'arrêt), ce qui écarta les curieux. Dans ce contexte, les SVM suscitèrent un engouement rapide et firent « oublier » les réseaux de neurones pour une dizaine d'années, jusqu'à l'émergence du *deep learning*, ou apprentissage profond, vers 2006.

1.3 Les SVM

Les « *Support Vector Machines* », ou SVM, permettent de résoudre de manière élégante les deux difficultés des perceptrons signalées précédemment : solutions multiples dans le cas de séparation parfaite et limitation au cas de la séparation linéaire. Les SVM combinent d'une part l'obtention d'une frontière de marge maximale (voir chapitre **??**) et d'autre part le plongement des données dans un espace plus vaste, que l'on peut appeler espace étendu qui est un « RKHS » ou espace de Hilbert à noyau reproduisant, afin de trouver des solutions non-linéaires. On doit à Antoine Cornuéjols (Cornuéjols et Miclet [2002]) la très astucieuse traduction de SVM par « Séparateur à Vaste Marge ».

La recherche de séparateurs à marge maximale dans le cas linéaire remonte à un article de Vladimir Vapnik et Alexandre Lerner de 1963 sur l'utilisation de la méthode des portraits généralisés en reconnaissance des formes : Vapnik [1963]. On en trouvera une présentation récente dans Chervonenkis [2013]. Le site `http://www.svms.org/history.html` de Martin Sewell mentionne ensuite les apports de Cover [1965], Mangasarian [1965], l'introduction de variables d'écart par Smith [1968] et la discussion de Duda et Hart [1973]. On notera avec intérêt que la première référence de M.Sewell est le célèbre article (Fisher [1936]) où R.A.Fisher introduisit son classifieur linéaire dans un problème de discrimination. Plus tard Krauth et Mézard [1987] publièrent l'algorithme « minover » qui attira l'attention d'Isabelle Guyon ; dans un récent billet elle (Guyon [2016]) revendique avoir effectué le lien entre les séparateurs linéaires

à vaste marge et l'usage des RKHS avec le « *kernel trick* ». L'article de Boser *et al.* [1992] présenté à la conférence COLT est en effet la première présentation des SVM sous leur forme désormais courante.

La théorie des espaces de Hilbert à noyaux reproduisants a été développée par Aronszajn [1950] ; Aizerman *et al.* [1961] donnèrent l'interprétation géométrique des noyaux comme produit scalaire dans l'espace étendu. On se reportera à Berlinet et Thomas-Agnan [2011] pour une monographie sur l'usage des RKHS en probabilités et statistique. Le « *kernel trick* » systématiquement utilisé consiste à munir l'espace étendu d'un produit scalaire fonction simple de celui de l'espace initial, ce qui évite d'avoir à effectuer des calculs dans un espace de grande dimension. Les noms de Vladimir Vapnik et Alexey Cervonenkis (ou Chervonenkis) sont indissociables de l'invention des SVM et du développement de la théorie statistique de l'apprentissage qu'ils fondèrent entre 1962 et 1973. On lira avec intérêt les deux livres d'hommage pour leurs 75 $^{\text{èmes}}$ anniversaires, consacrés l'un à Alexey Cervonenkis (Vovk *et al.* [2015]) et l'autre à Vladimir Vapnik (Luo *et al.* [2013]), ainsi que le numéro spécial du *Journal of Machine Learning Research* de septembre 2015 publié à la suite du décès de A.Cervonenkis.

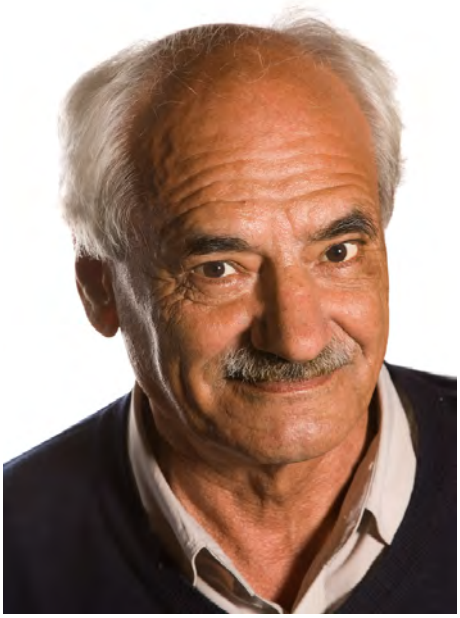

Figure 1.9 : *Alexey Cervonenkis.*

Alexey Cervonenkis naquit à Moscou en 1938 et étudia à partir de 1955 à la Faculté de radio ingénierie de l'Institut de Physique et Technologie de

Moscou. Après son doctorat en 1961 il obtint un poste à l'ICS où il fit toute sa carrière. Il partagea les 15 dernières années de sa vie entre l'Académie des Sciences de Russie et le département d'Informatique du Royal Holloway, un des sept collèges de l'université de Londres, où il fut nommé professeur en 2000. A. Cervonenkis décéda tragiquement d'hypothermie le 21 septembre 2014 : bien qu'excellent marcheur (il pouvait parcourir de 15 à 20 km par jour) il s'égara lors d'une promenade solitaire dans le parc national Lossiny Ostrov (l'ile aux élans) de Moscou, une forêt de plus de 100 km^2. Il téléphona à son épouse qui prévint les secours, mais les 182 sauveteurs ne purent le trouver et ce n'est que deux jours plus tard qu'à l'aide d'un hélicoptère on put localiser son corps.

Vladimir Vapnik est né en 1936 en Union Soviétique ; il obtint en 1958 un mastère de mathématiques à l'Université d'État d'Ouzbékistan, à Samarkand, puis en 1964 un doctorat de statistique à l'Institut des Sciences de Contrôle, à Moscou. Il travailla dans ce même institut de 1961 à 1990, et en devint le directeur du département de recherche en informatique. Il quitta l'Union Soviétique après 1990 pour les ATT-Bell Labs. Après avoir travaillé aux laboratoires NEC de Princeton ainsi qu'à l'université Columbia, il a rejoint en novembre 2014 le laboratoire de recherche sur l'intelligence artificielle de Facebook.

Figure 1.10 : *Vladimir Vapnik, Bechyné, République tchèque, 1990.*

Leur article fondamental Vapnik et Chervonenkis [1968] sur la convergence uniforme des fréquences d'occurrences d'événements vers leurs probabilités fut publié dans une des revues de l'académie des sciences soviétique, mais soumis deux ans auparavant. Pourquoi un tel délai peu fréquent à l'époque ? Gammerman et Vovk [2015] l'expliquent ainsi : les deux auteurs souhaitaient que l'article soit présenté par Andrei N. Kolmogorov, mais après une longue attente Boris Vladimirovitch Gnedenko leur expliqua que ce n'était pas de la statistique au sens que Kolmogorov et lui-même donnaient à ce mot et qu'il ne pouvait donc être publié sous cet intitulé. Finalement l'article fut publié sans changement dans la rubrique « cybernétique » et présenté par Vadim Trapezni-

kov, le directeur de l'Institute of Automation and Remote Control (dénommé ultérieurement Institute of Control Science ou ICS en anglais) dont dépendait leur laboratoire dirigé par Alexandre Lerner, alors en compétition avec celui de Aizerman, dans le même institut ! Novoseltsev [2015] rapporte que l'on disait que Vapnik inventait souvent ce que Cervonenkis prouvait ensuite.

L'ouvrage essentiel de la théorie statistique de l'apprentissage fut publié en russe en 1974 (Vapnik et Chervonenkis [1974]) puis traduit en 1979 en allemand par Akademie Verlag. En 1979 V.Vapnik publia en russe « *Estimation of Dependences Based on Empirical Data* », traduit en anglais en 1982 [Vapnik et Kotz, 1982]. La deuxième édition (Vapnik [2006]) publiée 25 ans après la première est enrichie d'une centaine de pages exposant la philosophie des sciences (*Empirical Inference Science*) sous-tendant les recherches de Vladimir Vapnik. Entretemps Vapnik a publié en 1995 *The Nature of Statistical Learning Theory* (Vapnik [1995]) ou « livre jaune » de 300 pages, puis en 1998 *Statistical Learning Theory* (Vapnik [1998]) le « livre gris » de près de 800 pages, qui font réference.

1.4 Sur quelques outils

1.4.1 La régression logistique

Sigmoïde, tangente hyperbolique, logit, régression logistique, etc. : sous différents noms la fonction logistique est utilisée pour estimer la probabilité d'un événement dans de nombreux domaines d'application. En apprentissage, on la trouve comme fonction d'activation d'un neurone, et comme classifieur.

L'article Cramer [2003b], qui est une version augmentée du chapitre 9 du livre du même auteur Cramer [2003a] fournit un historique très utile. L'origine de la formule

$$P(Z) = \frac{\exp(Z)}{1 + \exp(Z)}$$

remonte au XIX $^{\text{ème}}$siècle avec les travaux du mathématicien belge Pierre-François Verhulst (1804-1849), un élève d'Alphonse Quételet, sur les modèles de croissance non exponentielle des populations. Verhulst, inspiré par l' « Essai sur le principe de population » de Thomas Malthus, l'obtint à partir d'une équation différentielle et la baptise fonction logistique [1845] : Verhulst [1845]. Elle est redécouverte en 1920 par Pearl et Reed, toujours en démographie.

C'est ensuite au tour de la biologie, du fait du seul Joseph Berkson (1899-1982) que sous le nom de modèle logit, elle concurrence non sans oppositions le modèle probit. Berkson défendait l'estimation par le minimum du chi-deux plutôt que par le maximum de vraisemblance. Dans tous les cas, les calculs étaient difficiles et il fallut attendre 1977 pour disposer d'une procédure efficace dans le logiciel BMDP. Le débat idéologique s'apaisa dans les années 60 et le lien avec l'analyse discriminante reconnu dans le cas de la classification binaire avec

modèle gaussien : la probabilité *a posteriori* d'appartenir à l'une des classes est alors fonction logistique du score de Fisher. Cette propriété est à rapprocher de la formule de Platt *et al.* [1999] qui transforme en probabilité la valeur du classifieur obtenu par un SVM.

David Cox étudia l'extension au cas polytomique qui fut redécouverte par Theil [1969] avec qui vint l'ère des économistes. Daniel McFadden (McFadden [1973]) relia le modèle logit polytomique à la théorie du choix discret, ce qui lui valut le prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel, dit « prix Nobel d'économie » en 2000.

1.4.2 La validation croisée

Vérifier un modèle en le testant sur des données qui n'ont pas servi à le choisir ni à l'estimer est devenu une pratique indispensable en apprentissage, en tous cas pour l'apprentissage supervisé. Les techniques sont variées allant de la rotation de k sous-groupes au découpage répété des données en deux (apprentissage, test) ou trois (apprentissage, test, validation) parties. Plutôt que l'ajustement d'un modèle à des données à l'aide de critères de type chi-deux, Kolmogorov-Smirnov, ou encore vraisemblance pénalisée, on cherche à mesurer la capacité prédictive ou de généralisation sur des données indépendantes.

On attribue généralement cette pratique au *machine learning* qui l'a systématisé pour éviter les phénomènes de surajustement mais on en trouve les prémices dans les travaux de Lachenbruch et Mickey [1968] en analyse discriminante qui introduisent le *leave one out* pour estimer la performance d'une méthode, ainsi que dans Allen [1974] avec la statistique PRESS ou *predicted residual error sum of squares* en régression.

L'article de Stone [1974] cite les travaux du célèbre psychométricien Paul Horst (1903-1999) : « *Horst [1941], in his fascinating study of the prediction of success in marriage, found a "drop in predictability" between an "original" sample and a "check" sample that depended strongly on the method of construction of the predictor* ». Le chapitre X de cet ouvrage Horst *et al.* [1941], intitulé « Testing of prediction procedure and revision of hypotheses » énonce clairement « *the usefulness of a prediction procedure is not established when it is found to predict adequately on the original sample ; the necessary next step must be its application to at least a second group. Only if it predicts adequately on subsequent samples can the value of the procedure be regarded as established* ». Parcourir la littérature statistique et les syllabus universitaires montre que Horst ne fut guère suivi pendant des décennies !

1.5 Le débat entre modélisation et prévision

En 2001 Leo Breiman (1928-2005) ([Breiman *et al.*, 2001]) publie dans *Statistical Science* un article retentissant intitulé « *Statistical Modeling: The Two*

Cultures » . Le résumé donne le ton : « *There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools* ». Breiman se livre à une critique en règle des abus de la modélisation, en particulier dans des journaux réputés, pour lesquels il note « *The question of how well the model fits the data is of secondary importance compared to the construction of an ingenious stochastic model* ». Il défend l'idée d'exploiter des algorithmes d'apprentissage et reprend le paradigme de la boîte noire en des termes similaires à ceux de Vapnik :

Figure 1.11 : *Le modèle de la boîte noire.*

Si f est le modèle stochastique qu'utilise la boîte noire $y = f(x) + \varepsilon$, on cherche une fonction qui approxime en un certain sens le comportement de la boîte noire. Deux conceptions différentes s'opposent alors : soit on cherche à approximer la vraie fonction f, soit on cherche à obtenir des prévisions de y aussi précises que possible.

Dans la discussion qui suit l'article, David Cox, illustre représentant de la culture de la modélisation, exprime son profond désaccord, tandis qu'Emmanuel Parzen et Bruce Hoadley partagent les vues de Breiman. Leo Breiman n'était pas un statisticien ordinaire : il est certes bien connu comme l'un des pères de la méthode CART, du bagging, des forêts aléatoires (voir chapitre **??**) et du stacking, une méthode de combinaison de modèles, mais il eut une carrière non linéaire !

Après son doctorat en probabilités sur les processus homogènes, préparé sous la direction de Michel Loève à Berkeley en 1954, il obtint un poste à l'UCLA. Considérant que l'enseignement des mathématiques dans le secondaire était

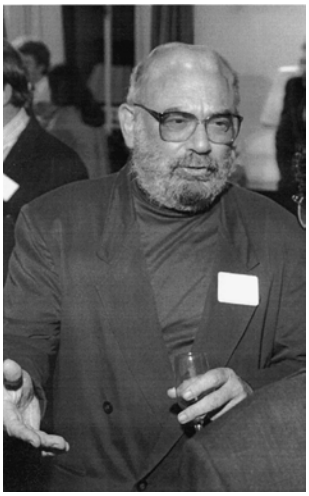

Figure 1.12 : *Leo Breiman en 1995.*

inadapté, il se porta volontaire pour enseigner dans le secondaire. Il prit une année sabbatique pour estimer le nombre d'élèves au Libéria alors qu'il n'y avait quasiment pas de routes. Il démissionna de l'UCLA après sept ans d'exercice pour éviter de devenir un mathématicien abstrait et devint consultant pendant 13 ans. Il reprit ensuite un poste à Berkeley où il resta jusqu'à sa retraite en 1993, puis comme professeur émérite sans s'arrêter de produire. Il avait également des talents de sculpteur et d'architecte. Pour en savoir plus, on se reportera avec profit à son entretien avec Richard Olshen : Olshen [2001].

Très récemment David Donoho (Donoho [2015]) a repris la discussion entamée par Breiman. En voici quelques extraits : « *The generative modelling culture seeks to develop stochastic models which fits the data, and then make inferences about the data-generating mechanism based on the structure of those models. Implicit (?) is the notion that there is a true model generating the data, and often a truly 'best' way to analyze the data. The predictive modelling culture is silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets. Machine Learning is identified by Breiman as the epicenter of the Predictive Modeling culture.* »

L'opposition entre comprendre et prévoir a été également discutée dans Saporta [2008] et Shmueli [2010]. Le fait de pouvoir prévoir sans chercher à comprendre le mécanisme générateur des données peut en effet choquer un esprit scientifique, et il y a bien une ligne de rupture entre les tenants de la modélisation avec des modèles parcimonieux et interprétables, et les partisans de la science empirique qui engrangent des succès dans le monde du traitement

de données industrielles massives.

Faut-il opposer les deux approches apprentissage et modélisation, ou au contraire voir leurs complémentarités comme le suggère Hal Varian aux économistes (Varian [2014]) ? Les chapitres suivants illustreront le deuxième point de vue, rendu indispensable par l'explosion des données.

Bibliographie

Aizerman, M., Braverman, E. et Rozonoer, L. [1961]. Method of potential functions in the problem of restoration of a functional converter characteristic by means of points observed randomly. *Avtomatika i Telemekhanika*, **25**(12).

Allen, D. M. [1974]. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**(1), 125–127.

Aronszajn, N. [1950]. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, **68**(3), 337–404.

Berlinet, A. et Thomas-Agnan, C. [2011]. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer.

Boser, B. E., Guyon, I. M. et Vapnik, V. N. [1992]. A training algorithm for optimal margin classifiers. Dans *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152. ACM.

Bottou, L. Y. [1991]. *Une approche théorique de l'apprentissage connexionniste ; applications à la reconnaissance de la parole*. Thèse de doctorat, Université Paris XI.

Breiman, L. *et al.* [2001]. Statistical modeling : The two cultures (with comments and a rejoinder by the author). *Statistical Science*, **16**(3), 199–231.

Chervonenkis, A. Y. [2013]. Early history of support vector machines. Dans *Empirical Inference*, 13–20. Springer.

Conway, F., Siegelman, J. et Alexanderson, G. L. [2005]. *Dark hero of the information age : In search of Norbert Wiener, the father of cybernetics*. Basic Books.

Conway, F., Siegelman, J., Vallée, N. et Vallée, R. [2012]. *Héros pathétique de l'âge de l'information : en quête de Norbert Wiener, père de la cybernétique*. Hermann.

Copeland, B. J. et Proudfoot, D. [1996]. On alan turing's anticipation of connectionism. *Synthese*, **108**(3), 361–377.

Cornuéjols, A. et Miclet, L. [2002]. *Apprentissage Artificiel : Concepts et Méthodes*. Eyrolles.

Cover, T. M. [1965]. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, 326–334.

Cramer, J. S. [2003a]. *Logit models from economics and other fields*. Cambridge University Press.

Cramer, J. S. [2003b]. The origins and development of the logit model. https ://www.cambridge.org/resources/0521815886/1208_default.pdf.

Cybenko, G. [1989]. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**(4), 303–314.

Donoho, D. [2015]. 50 years of data science. Dans *Princeton NJ, Tukey Centennial Workshop*.

Duda, R. et Hart, P. [1973]. *Pattern classification and scene analysis*. Wiley.

Fisher, R. A. [1936]. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, **7**(2), 179–188.

Fogelman-Soulie, F., Robert, Y. et Tchuente, M. [1987]. *Automata networks in computer science : theory and applications*. Manchester University Press .

Franklin, C. et Agresti, A. [2013]. *Statistics : The art and science of learning from data*. Pearson Prentice Hall.

Gammerman, A. et Vovk, V. [2015]. Alexey chervonenkis's bibliography : Introductory comments. *Journal of Machine Learning Research*, **16**, 2051–2066.

Guyon, I. [2016]. Data mining history : The invention of support vector machines. http ://www.kdnuggets.com/2016/07/guyon-data-mining-history-svm-support-vector-machines.html.

Hastie, T., Tibshirani, R. et Friedman, J. [2009]. *The elements of statistical learning : data mining, inference, and prediction*. Springer. Second edition.

Horst, P., Wallin, P. C., Guttman, L. C., Wallin, F. B. C., Clausen, J. A., Reed, R. C. et Rosenthal, E. C. [1941]. *The prediction of personal adjustment : A survey of logical problems and research techniques, with illustrative application to problems of vocational selection, school success, marriage, and crime*. Social science research council.

Kohonen, T. [1989]. *Self-Organization and Associative Memory*. Springer, third édition.

Krauth, W. et Mézard, M. [1987]. Learning algorithms with optimal stability in neural networks. *Journal of Physics A : Mathematical and General*, **20** (11), L745.

Lachenbruch, P. A. et Mickey, M. R. [1968]. Estimation of error rates in discriminant analysis. *Technometrics*, **10**(1), 1–11.

Luo, Z., Schölkopf, B. et Vovk, V. [2013]. *Empirical Inference : Festschrift in Honor of Vladimir N. Vapnik*. Springer.

Mangasarian, O. L. [1965]. Linear and nonlinear separation of patterns by linear programming. *Operations research*, **13**(3), 444–452.

McCulloch, W. S. [1974]. Recollections of the many sources of cybernetics. *ASC Forum*, **6**(2), 5–16.

McCulloch, W. S. et Pitts, W. [1943]. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**(4), 115–133.

McFadden, D. [1973]. Conditional logit analysis of qualitative choice behavior. Dans *Frontiers in Econometrics*, 105–142. Academic Press.

Minsky, M. et Papert, S. [1969]. *Perceptrons : an Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA.

Novoseltsev, V. N. [2015]. Institute of control sciences through the lens of vc dimension. Dans *Measures of Complexity*, 43–53. Springer.

Olshen, R. [2001]. A conversation with leo breiman. *Statistical Science*, 184–198.

Platt, J. *et al.* [1999]. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, **10**(3), 61–74.

Rosenblatt, F. [1962]. *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, DC, USA.

Saporta, G. [2008]. Models for understanding versus models for prediction. Dans *COMPSTAT 2008*, 315–322. Springer.

Shmueli, G. [2010]. To explain or to predict ? *Statistical science*, 289–310.

Smith, F. W. [1968]. Pattern classifier design by linear programming. *IEEE Transactions on Computers*, **100**(4), 367–372.

Stone, M. [1974]. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 111–147.

Theil, H. [1969]. A multinomial extension of the linear logit model. *International Economic Review*, **10**(3), 251–59.

Turk, M. et Pentland, A. [1991]. Eigenfaces for recognition. *Journal of cognitive neuroscience*, **3**(1), 71–86.

Vapnik, V. [1963]. Pattern recognition using generalized portrait method. *Automation and remote control*, **24**, 774–780.

Vapnik, V. [1995]. *The nature of statistical learning theory*. Springer.

Vapnik, V. [1998]. *Statistical Learning Theory*. Wiley.

Vapnik, V. [2006]. *Estimation of dependences based on empirical data, second edition*. Springer.

Vapnik, V. N. et Chervonenkis, A. Y. [1968]. On the uniform convergence of relative frequencies of events to their probabilities. Dans *Soviet Math. Dokl*, volume 9, 915–918.

Vapnik, V. N. et Kotz, S. [1982]. *Estimation of dependences based on empirical data*. Springer.

Vapnik, V. et Chervonenkis, A. Y. [1974]. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya (Theory of pattern recognition. Statistical problems of learning)*. Moscow : Nauka.

Varian, H. R. [2014]. Big data : New tricks for econometrics. *The Journal of Economic Perspectives*, **28**(2), 3–27.

Verhulst, P. F. [1845]. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, **18**, 14–54.

Von Neumann, J. [1945]. First draft of a report on the edva. between the united states army ordinance department and the university of pennsylvania moore school of electrical engineering university of pennsylvania. *Contract No. W-670-ORD-4926*.

Vovk, V., Papadopoulos, H. et Gammerman, A. [2015]. *Measures of Complexity*. Springer.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. et Lang, K. J. [1989]. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, **37**(3), 328–339.

Wiener, N. [1948]. *Cybernetics : Control and communication in the animal and the machine*. Hermann et MIT Press.