Chapitre 1. Introduction à l'apprentissage statistique

Claude Petit, Insee et université de Rennes - claude.petit@univ-rennes.fr

Oct. 2025

Overview

1. Définition du Machine Learning

Définition 1/2 : apprendre à partir des données

Beaucoup de noms différents : apprentissage statistique, apprentissage automatique, Machine Learning (ML), Statistical Learning.

Wikipedia : champ d'études de l'intelligence artificielle (IA) qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'apprendre à partir des données.

ChatGPT : c'est une sous-discipline de l'IA qui permet aux ordinateurs d'apprendre à partir de données et de prendre des décisions sans être explicitement programmés pour accomplir ces tâches.

- Imiter (un peu) le comportement inductif du cerveau pour prendre des décisions de façon autonome, en fonction des données.
- Domaine à l'intersection de l'informatique et des mathématiques.
- Différence avec les statistiques via l'outil de base: le modèle en statistique et l'algorithme en ML.

En bref: le ML, ce sont les mathématiques de l'IA.

Définition 2/2 : c'est l'une des branches de l'IA

Fausses branches de l'IA : apprentissage automatique, traitement du langage naturel, vision par ordinateur et robotique.

En fait, tout cela fait partie du ML.

Vraies branches de l'IA :

- Apprentissage statistique (approche connexionniste) : ML, réseaux de neurones, apprentissage par renforcement, traitement du langage naturel.
- Systèmes formels (approche cognitive) : programmation logique, machines de Turing, calculabilité, théorie des langages (Chomsky).
- Méthodes faibles (approche pragmatique) : heuristiques, problèmes de satisfaction de contraintes, systèmes experts, représentation de connaissances.

Sous domaines de l’IA

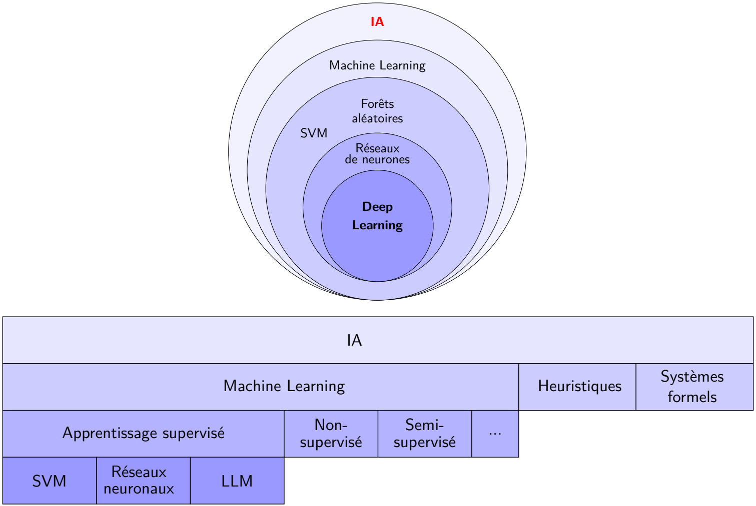


Figure 1: Les différents sous domaines de l’IA.

Les différentes approches du ML

L'apprentissage statistique peut être :

- Supervisé : données étiquetées. Deviner les étiquettes des nouvelles données. Phase d'entrainement, puis de test. Ex : classification et régression, kNN, arbres de décisions, SVM, réseaux de neurones.
- Non supervisé : données non étiquetées. Caractériser la loi de proba ayant engendré ces observations. Ex : algorithmes de clustering, estimation de densités, classifications hiérarchiques, DBScan, MeanShift, ACP, ACM.
- Semi-supervisé : petite partie des données étiquetée. Ex : GNNs (Graph Neural Networks), modèles génératifs bayésiens, TSVM, régularisation de laplaciens.
- Par renforcement : Q-learning, SARSA, gradient de politique, réseaux de neurones antagonistes (GAN).
- En ligne : perceptron, méthode de descente de gradient stochastique (SGD), filtres de Kalman, algorithmes de Bandits.

Quelques dates

• le ML date des années 50 (conf. de Darthmouth, été 1956), mais deux étapes très importantes ensuite :

- Années 1980 : réseaux de neurones et algorithmes de rétropropagation de gradient.
- Années 2000 : lien avec les statistiques par Vapnik (the nature of Statistical Learning).

Les hivers de l'IA (« Winter is coming ») :

- 1973-1980. Limite des perceptrons monocouches, manque d'avancées en robotique et traitement automatique du langage (TAL).
- 1987-1993. Échec commercial des machines LISP, abandon des systèmes experts, réseaux de neurones peu efficaces.

Quelques ouvrages

- The Elements of Statistical Learning, Hastie, Tibshirani, Friedman, éd. Springer (trop gros).
- A probabilistic Theory of Pattern Recognition, Devroye, Lugosi (trop gros).
- An Introduction to Statistical Learning with Application in Python, James, Witten, Hastie, Tibshirani, Taylor, éd. Springer.
- Introduction à l'apprentissage automatique, de Frédéric Sur, polycopié de l'école des Mines de Nancy (très bien).
- Mathematics for Machine Learning, de G. Thomzd, polycopié de l'université de Berkeley.
- Algorithme, la bombe à retardement (Weapons of Math Destruction), de Cathy O'Neil, éd. Les Arènes (trop bien).
- Contre-atlas de l'intelligence artificielle, de Kate Crawford, éd. Zulma (trop bien).
- La guerre des intelligences, Laurent Alexandre, éd. J.C. Lattès (à lire pour le critiquer en toute connaissance de cause).

2. Le problème de la dimension

La réduction de dimension

• Réduction de dimension : transformer des données d'un espace de grande dimension en un espace de dimension inférieure tout en préservant des propriétés essentielles des données d'origine.

• Objectif : rendre possible ou plus rapide le traitement de ces données, réduire la complexité des processus, économiser de l'espace, de l'énergie, du temps, se prémunir contre le fléau de la dimension, améliorer l'interprétabilité, visualiser des données.

• Méthodes : linéaires ou non linéaires, aléatoires ou déterministes.

Notre classification découle du paradigme du big data : 2 paramètres fondamentaux décrivent les dimensions des données : n, taille de la population (nombre d'éléments de la base de données) et d, dimension des variables statistiques attachées à ces éléments.

Les paramètres fondamentaux de la taille en statistique

Trois situations possibles :

- n grand, d petit : domaine des statistiques multivariées traditionnelles (analyse de données « à la française »). Les outils d'inférence statistique classiques fonctionnent bien, théorèmes limites classiques (n tend vers l'infini avec d fixé).
- n petit, d grand : domaine des statistiques en grande dimension. Les outils statistique usuels ne fonctionnent plus. Matrice de covariance empirique singulière, estimateurs des moindres carrés non consistants, etc. Hypothèses suppl. nécessaires pour traiter les données : parcimonie, structure sous-jacente cachée ayant une petite dimension, etc.
- n et d grands : autre aspect des statistiques en grande dimension, domaine des matrices aléatoires. Aucun théorème limite classique ne s'applique, hypothèses sur limite de n/d quand n et d tendent vers l'infini, nécessaires pour appliquer des théorèmes spécifiques.

Le fléau de la dimension -1

- « The Curse of Dimensionality » : expression de Richard Bellman 1950-1960.
- Problèmes en apprentissage dus aux propriétés des espaces de grande dimension.
- Espaces en grande dimension : l'intuition par rapport à la dimension 1,2 ou 3 est parfois fausse.
- Lié au dilemne biais / variance.

Premier exemple : si d décision binaires doivent être effectuées pour optimiser une fonction de perte, la recherche exhaustive de l'optimum nécessite 2^d évaluations, qui augmente de façon exponentielle avec d.

Le fléau de la dimension -2 : taille des échantillons

Second exemple : On partitionne le cube unité de $\mathbb{R}^d$ en cubes de côté $1/n$. n^d nécessaires pour remplir le cube unité. Estimer loi de proba. à partir d'un échantillon de taille n : précision $1/10$ en dimension 1 (une mesure par petit cube). Même finesse en dimension $d = 10$: échantillon de taille $n = 10^{10}$ nécessaire.

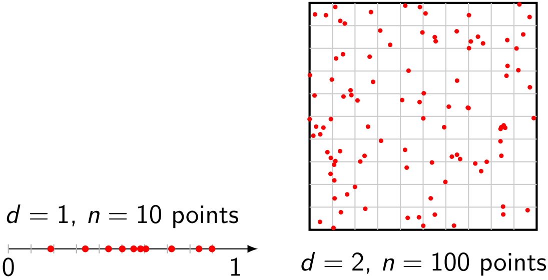


Figure 2: Partition du cube unité et échantillonnage en dimension $d = 1, 2, 3$. Taille de grille $= 1/10$. Le nombre de points nécessaires explose avec la dimension.

Le fléau de la dimension -2 : taille des échantillons

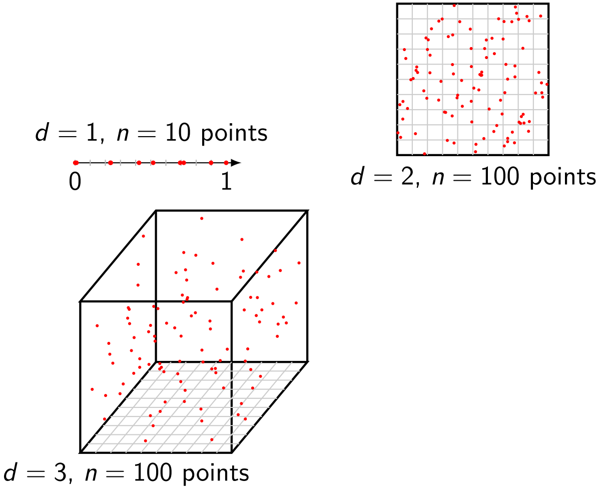


Figure 3: Partition du cube unité et échantillonnage en dimension $d = 1, 2, 3$. Taille de grille $= 1/10$. Le nombre de points nécessaires explose avec la dimension.

Le fléau de la dimension -3 : volume de la boule unité

Troisième exemple : volume de la boule unité en dim. d :

$$V_d = \pi^{d/2}\Gamma(d/2+1)^{-1}$$

Γ fonction Gamma d'Euler. En utilisant la formule de Stirling, $\Gamma(x) \sim \sqrt{2\pi}x^{x-1/2}e^{-x}$, on a :

$$V_d \sim \frac{1}{\sqrt{\pi d}}\left(\frac{2\pi e}{d}\right)^{d/2} \to 0 \text{ si } d \to +\infty$$

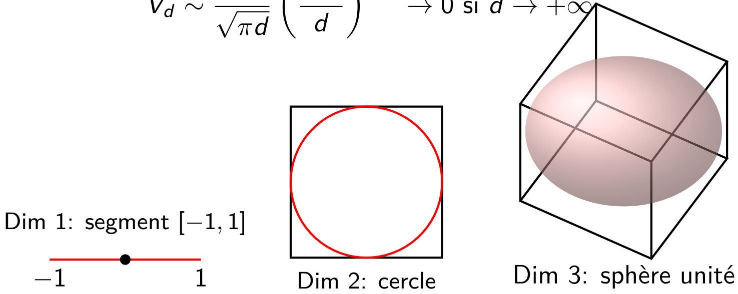


Figure 4: Volume de la boule unité en dimension $d = 1, 2, 3$.

Le fléau de la dimension -3 : hypercubes

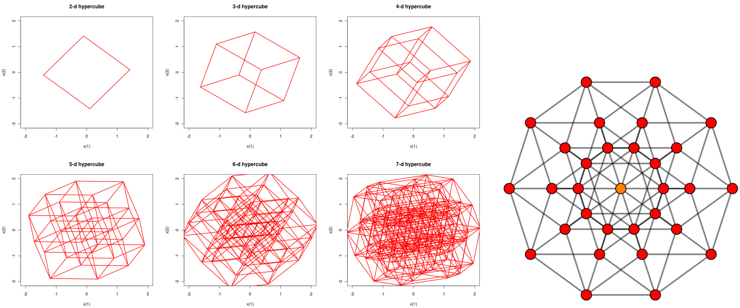


Figure 5: À gauche : hypercubes pour d=2,3,4,5,6,7. À droite : représentation « à plat » du graphe correspondant à l'hypercube de dimension 5. Crédit : Tom Ruan, Wikipédia.

Le fléau de la dimension -3 : hypercubes

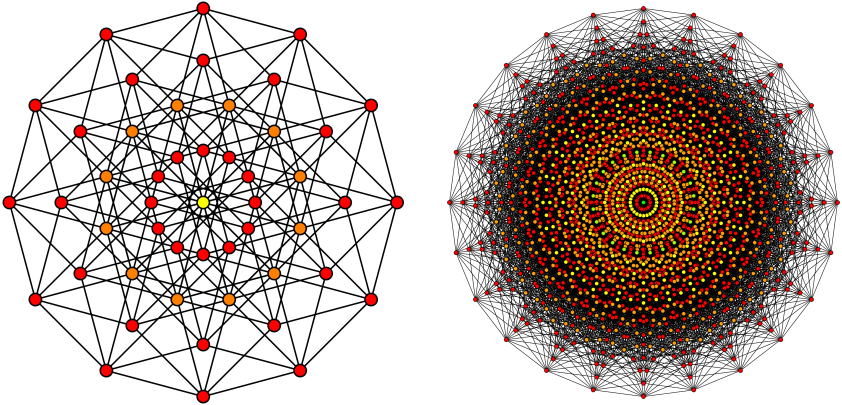
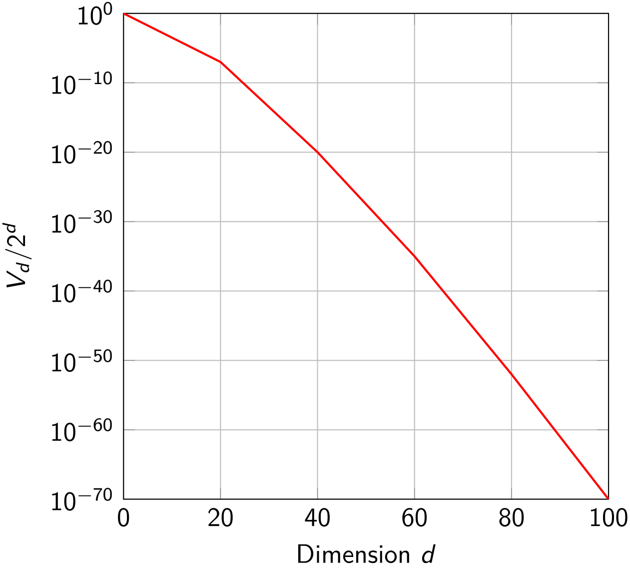


Figure 6: hypercubes « à plat » pour d=6,9. Crédit : Tom Ruan, Wikipédia.

Le fléau de la dimension -3 : le ratio $V_d/2^d$

Le rapport $V_d/2^d$ tend vers 0 encore plus vite !



$(1-\epsilon)^d V_d$ volume de la boule de rayon $1-\epsilon$ et

$$\frac{V_d - (1-\epsilon)^d V_d}{V_d} = 1 - (1-\epsilon)^d$$

proportion de la boule unité dans une couche d'épaisseur $\epsilon > 0$ à la surface de la boule. $\longrightarrow 1$ si $d \to +\infty$: en grande dim., tout le volume est concentré à la surface.

Conséquence : des points choisis aléatoirement dans l'espace se concentrent à la surface de la boule $\Rightarrow$ la distance n'est plus pertinente en grande dim. Tous les points deviennent équidistants.

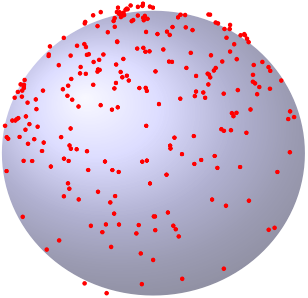


Figure 7: Points aléatoires générés de façon uniforme dans l'espace, se concentrant à la surface de la sphère.

Le fléau de la dimension - 4 : gaussiennes en grande dimension

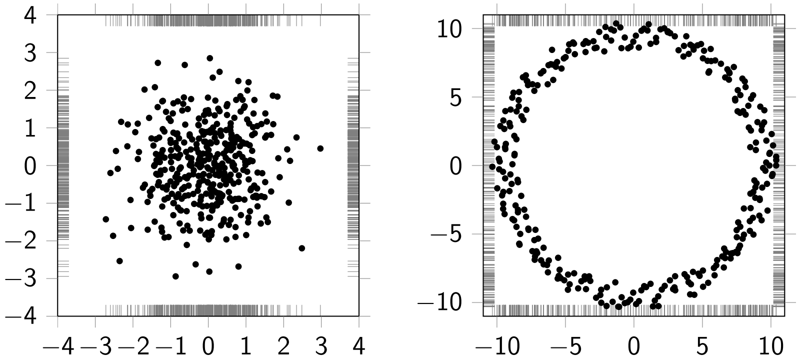


Figure 8: Gauche : échantillon gaussien standard en dimension $d = 2$. Droit : Échantillon gaussien standard en dimension $d = 100$, projeté dans un sous-espace de dimension 2.

Le fléau de la dimension 5 : le phénomène de Hughes

À base d'apprentissage de taille fixée, les performances d'un classifieur augmentent avec la dimension, atteignent un maximum, puis diminuent.

• Heuristique publiée par Hughes en 1968 : il existe un nombre de « features » optimal à ne pas dépasser.

• En 1978, Van Campenhout montre que l'article contient des erreurs. Le paradoxe demeure (« On the Peaking of the Hughes Mean Recognition Accuracy: The Resolution of an Apparent Paradox », J.V. Van Campenhout. IEEE Transactions on Systems, 1978).

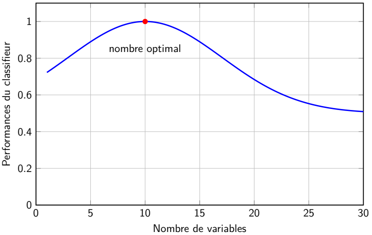


Figure 9: Illustration du paradoxe de Hughes

Solutions pour contrer la malédiction de la dimension

- Compresser les données.
- Trouver un sous-espace de petite dimension qui contient (presque) toute l'information.
- Réduire la dimension.

La malédiction de la dimension est liée au dilemne biais / variance et au problème de sous et sur apprentissage. Nous verrons cela dès le prochain chapitre.