RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Inria

Comment faire apprendre à une machine ?

Mathurin Massias

https://mathurinm.github.io

Inria, OCKHAM team

Séminaire SIESTE, 18/12/2024

Parlons de moi

- prépa + école d'ingénieur généraliste 2009 – 2014
- césure : appris le code pro (C# ...)
- M2 recherche Mathématiques Vision Apprentissage (+++)
- 1 an CDI chez Cardiologs (Deep learning/cardiologie)

Dans la recherche :

- thèse Télécom Paris + Inria Saclay 2016 – 2019
- postdoc Université de Gênes
- chargé de recherche Inria Lyon depuis nov. 2021
- prof. chargé de cours à Polytechnique, cours M1/M2 ENS
- ce que j'y trouve : liberté, stimulation, transmission, collaborations...

Outline

Comment en est-on arrivés là ?

Comment ça marche ?

Ma recherche

"selon une intelligence artificielle"

Tous | Actualités | Images | Vidéos | Livres | Web | Finance

À la une

Sud Ouest

Miss France 2025 : voici la gagnante du concours selon une intelligence artificielle

Il y a 2 jours

La Libre.be

Miss France 2025: voici qui succédera à Eve Gilles, selon une intelligence artificielle

il y a 19 heures

Dailymotion · Purepeople

L'identité de Miss France 2025 déjà connue ? La prédiction d'une Intelligenc…

VIDÉO Il y a 1 jour

entrevue.fr

Miss France 2025 : Miss Guadeloupe gagnante selon une intelligence artificielle…

il y a 12 heures

Voir plus →

Succès marquants

- mars 2016 : AlphaGo
- août 2017 : DeepL
- juillet 2020 : AlphaFold (Nobel chimie 2024)
- oct 2021 : Github Copilot
- janvier 2021 : Dall-E
- nov 2022 : ChatGPT

Raisons derrière ces avancées ?

Four major influences act today :

- *The formal theories of statistics*
- *Accelerating developments in computers and display devices*
- *The challenge, in many fields, of more and ever larger bodies of data*
- *The emphasis on quantification in an ever wider variety of disciplines*

Raisons derrière ces avancées ?

Four major influences act today :

- *The formal theories of statistics*
- *Accelerating developments in computers and display devices*
- *The challenge, in many fields, of more and ever larger bodies of data*
- *The emphasis on quantification in an ever wider variety of disciplines*

John Tukey, 1962

3 piliers : données, algos/modèles, calcul

Known for

- Exploratory data analysis
- Multiple comparisons problem
- Projection pursuit
- Box plot
- Blackman–Tukey transformation
- Cooley–Tukey FFT algorithm
- Freeman–Tukey transformation
- Siegel–Tukey test
- Stone–Tukey theorem
- Tukey–Duckworth test
- Tukey's range test
- Tukey lambda distribution
- Tukey's trimean
- Tukey's test of additivity
- Tukey's lemma
- Tukey mean difference plot
- Tukey median
- Tukey depth
- Tukey's biweight function
- Tukey's fences
- Tukey window
- Cepstrum
- Flexagon
- Median polish
- Midhinge
- Slash distribution
- Theory of conjoint measurement
- Coining the term 'bit'
- Scagnostics

Outline

Comment en est-on arrivés là ?

Comment ça marche ?

Ma recherche

Quelques tâches de Machine Learning

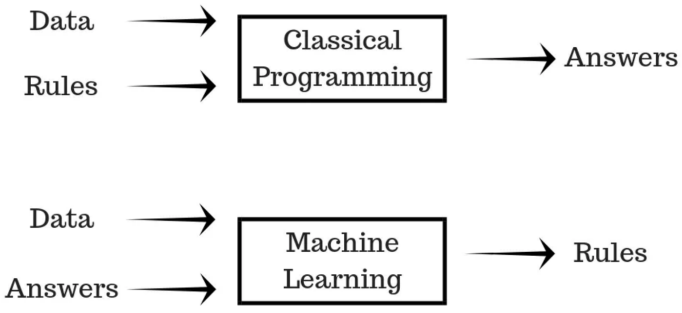
Modèle de ML : prend une entrée x, renvoie une sortie y (*prédiction*)

Exemples :

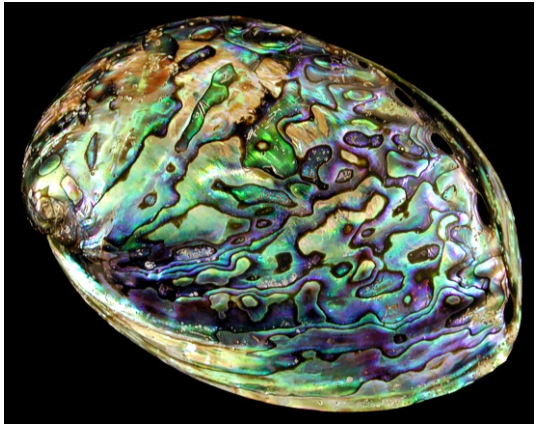
- $x =$ une photo $\to y =$ il y a un chat dessus ? $(0/1)$
- $x =$ position de pions sur un plateau de go $\to y =$ le *meilleur* mouvement à effectuer ?
- $x =$ un électrocardiogramme :
  - $y =$ normal/pathologique ? (0/1)
  - $y =$ liste des pathologies présentes ?
- $x =$ un texte en anglais $\to y =$ sa traduction française
- $x =$ une légende (texte) $\to y =$ une image correspondante ?
- $x =$ des caractéristiques d'une participante Miss France 2024 $\to y$ sa "proba" de gagner?

variété des types de données, variété des applications, mais cadre maths commun !

Différence avec le cadre “classique”



Passion mollusque



Tâche imaginaire : prédire l'âge d'un ormeau à partir de son poids et de son diamètre

Comment faire du ML ? Données

1) Besoin d'acquérir des **données** étiquetées

- vraies entrées x_i avec la bonne réponse y_i (*supervision*)
- beaucoup (différence notable avec l'humain)
- coûteux

1 bis) Chaque entrée : transformée en un vecteur $x_i \in \mathbb{R}^d$

- ormeau : $x_i =$ (poids, diamètre) $\in \mathbb{R}^2$
- image, signal audio : OK (a priori)
- texte : ?

$\hookrightarrow$ Collection $\{(\underbrace{(\text{poids 1}, \text{diamètre 1})}_{x_1 \in \mathbb{R}^2}, \underbrace{\text{âge 1}}_{y_1 \in \mathbb{R}}), (\underbrace{(\text{poids 2}, \text{diamètre 2})}_{x_2 \in \mathbb{R}^2}, \underbrace{\text{âge 2}}_{y_2 \in \mathbb{R}}), ...\} =$ données d'apprentissage

Comment faire du ML ? Modèle

2) On postule une forme pour la relation entrée/sortie : **choix du modèle**

Fonction paramétrique, par ex :

$$y = \text{age(ormeau)} = \theta_1 \times \text{poids} + \theta_2 \times \text{diamètre} = \theta^\top x = F_\theta(x)$$

Un modèle de ML, c'est une fonction paramétrique

Comment faire du ML ? Calcul

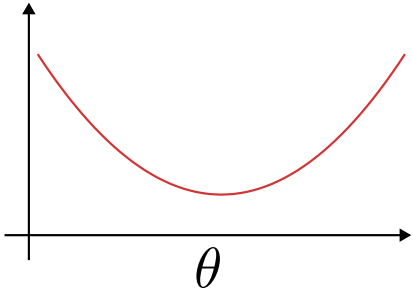
3) On optimise les paramètres du modèle: on trouve la valeur qui *marche le mieux*

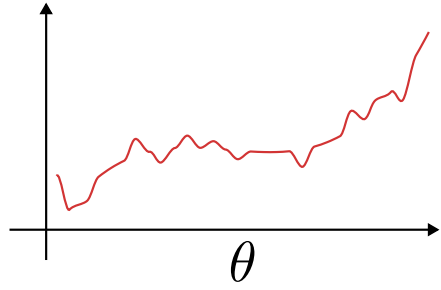
- mesure d'erreur entre prédiction $\hat{y}$ et vraie réponse y. Ex: $(\hat{y} - y)^2$
- erreur sur le i$^{\text{ème}}$ ormeau: $(y_i - x_i^\top \theta)^2$
- trouver le θ qui donne la plus petite erreur cumulée sur tous les ormeaux :

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} (y_i - x_i^\top \theta)^2$$

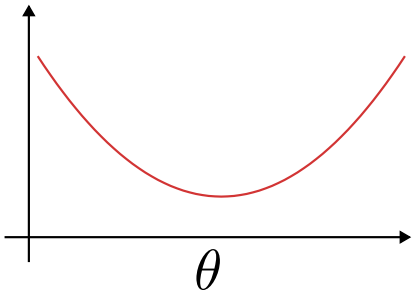
Entraîner un modèle de ML, c'est minimiser une fonction de coût

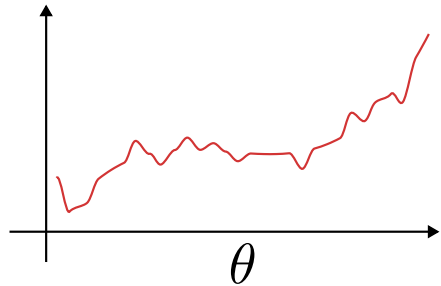
Comment minimiser des fonctions ?





Comment minimiser des fonctions ?

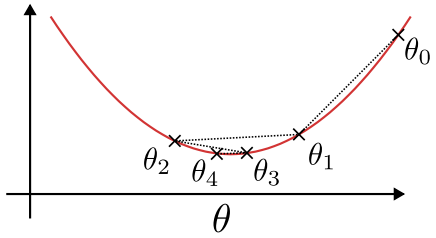




Importance de la convexité

Perdu dans la montagne : la descente de gradient

Perdu dans la montagne : la descente de gradient



Theorème : pour une fonction convexe la descente de gradient obtient une erreur de l'ordre de $1/k$ après k itérations :

$$\text{loss}(\theta_k) - \text{loss}(\theta^*) \leq \frac{\text{cste}}{k}$$

Défis en optimisation :

- nouveaux algorithmes, meilleures vitesses de convergence (algèbre linéaire, analyse convexe)
- implém efficace ! k vs temps effectif (algos distribués, GPU...)

Récap

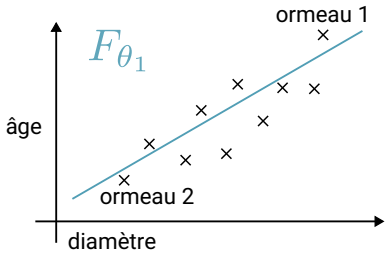
- acquérir des données (entrées/sorties)
- choisir un type de modèle (fonction paramétrique)
- l'entraîner (trouver les paramètres qui marchent le mieux)
- on a θ^*

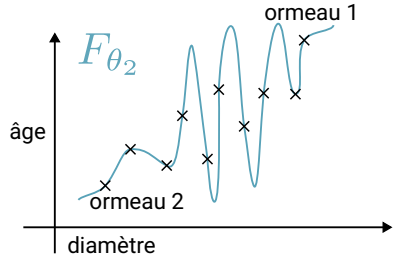
Récap

- acquérir des données (entrées/sorties)
- choisir un type de modèle (fonction paramétrique)
- l'entraîner (trouver les paramètres qui marchent le mieux)
- on a θ^*

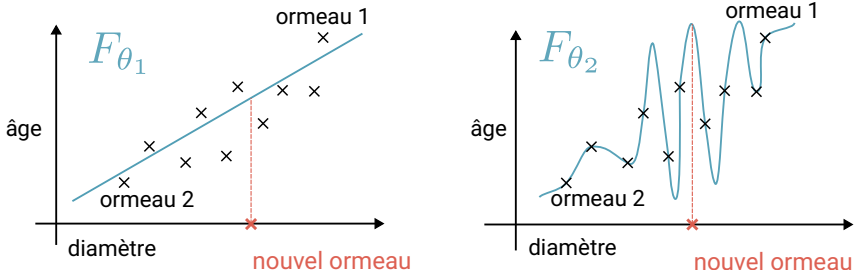
C'est si simple que ça ?

Qui est le meilleur modèle ?





Qui est le meilleur modèle ?



↪ Overfitting (surapprentissage), besoin de pénaliser les modèles trop complexes (?)

Défis : comprendre la **généralisation** des modèles (probas/stats)

...

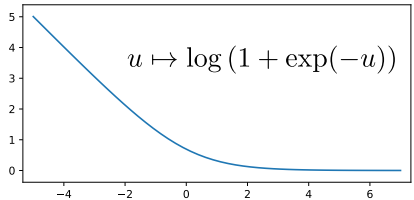
Classement final

Résultats	Candidates	Concours internationaux
Miss France 2025	• Miss Martinique — Angélique Angarni-Filopon	
1re dauphine	• Miss Nord-Pas-de-Calais — Sabah Aïb	
2e dauphine	• Miss Corse — Stella Vangioni	
3e dauphine	• Miss Guadeloupe — Moïra André	

Problèmes de classification

- Si $y \in \{-1, 1\}$ (e.g. distinguer ormeau *méditerranéen* vs ormeau *oreille d'âne*) ?
- Utiliser comme fonction/modèle $y = \text{signe}(\theta^\top x)$
- Fonction de coût différente :

$$\sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^\top x_i))$$



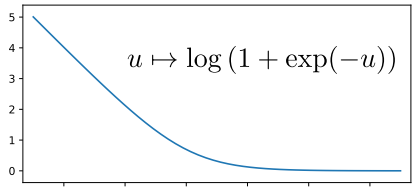
$\hookrightarrow$ pousse $y_i \theta^\top x_i$ à être grand donc $\theta^\top x_i$ à être du signe de y_i

- Et si j'ai $K > 2$ *classes*?

Problèmes de classification

- Si $y \in \{-1, 1\}$ (e.g. distinguer ormeau *méditerranéen* vs ormeau *oreille d'âne*) ?
- Utiliser comme fonction/modèle $y = \text{signe}(\theta^\top x)$
- Fonction de coût différente :

$$\sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^\top x_i))$$



$\hookrightarrow$ pousse $y_i \theta^\top x_i$ à être grand donc $\theta^\top x_i$ à être du signe de y_i

- Et si j'ai $K > 2$ *classes*?

$\text{softmax(z)} = \frac{\exp(z_k)}{\sum_{\ell=1}^{k} \exp(z_\ell)}$ me donne un vecteur dans $\mathbb{R}_+^K$ qui somme à 1

$$Y = \text{softmax}(\Theta x), \quad \Theta \in \mathbb{R}^{K \times d}$$

Deep learning

- une seule couche (perceptron) :

$$x \in \mathbb{R}^d \to \sigma(W_1 x + b_1)$$

  σ fonction non linéaire, typiquement ReLU = $\max(u, 0)$

- on enchaîne les *couches* :

$$x \to \sigma(W_1 x + b_1) \to \sigma(W_2 \sigma(W_1 x + b_1) + b_2) \to \ldots$$

- très flexible : peut approcher n'importe quelle fonction
- plein de paramètres (weights W_ℓ, biases b_ℓ)
- pas convexe du tout
- autres architectures (notamment les Transformers)

ChatGPT en 2 slides très approximatives

Next word prediction :

$$F(\text{Peux tu m'aider à faire mon DM ?}) = \text{Bien}$$

$$F(\text{Peux tu m'aider à faire mon DM ? Bien}) = \text{sûr,}$$

$$F(\text{Peux tu m'aider à faire mon DM ? Bien sûr,}) = \ldots$$

ChatGPT en 2 slides très approximative

- travaille avec un vocabulaire de taille n_{mots} (fixé)
- chaque mot envoyé vers un vecteur dans $\mathbb{R}^{10000}$
- fonction prend en entrée N mots consécutifs
- prédit un vecteur de probas dans $\mathbb{R}^{n_{\text{mots}}}$
- mot prédit = mot avec la plus grande proba
- on rajoute ce mot aux $N-1$ derniers mots et on recommence

- entraîné sur $\approx$ tout le texte d'Internet

Défis en deep learning

- besoin d'énormément de données
- énormément de paramètres: 176 milliards pour GPT3
- entraînement long
- beaucoup de ressources, machines dédiées (GPU)

Les théoriciens ont dit pendant des années que ça ne pouvait pas marcher :

- modèle trop complexes (overfittent)
- optimisation non convexe (trop difficile)

A part le supervisé

Supervisé : coûteux (e.g. médical, texte de qualité)

Pistes d'amélioration actuelles :

- transfert/fine-tuning (LLM)
- faible supervision
- autosupervision

Apprentissage par renforcement : jeu/simulation/exploration

Autres défis

- compréhension des modèles (pourquoi ça marche ?)
- frugalité
- interprétabilité (boîte noire ++)
- confidentialité des données d'entraînement
- fairness, biais
- open source, éducation, démocratisation

The bitter lesson ? by Rich Sutton

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. [...] Researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. [...]

[examples : chess, go, speech and image recognition]

One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. [...]

The second general point to be learned from the bitter lesson is that the actual contents of minds are tremendously, irredeemably complex; we should stop trying to find simple ways to think about the contents of minds, such as simple ways to think about space, objects, multiple agents, or symmetries. All these are part of the arbitrary, intrinsically-complex, outside world. They are not what should be built in, as their complexity is endless; instead we should build in only the meta-methods that can find and capture this arbitrary complexity. Essential to these methods is that they can find good approximations [...] We want AI agents that can discover like we can, not which contain what we have discovered.

Outline

Comment en est-on arrivés là ?

Comment ça marche ?

Ma recherche

Mes thèmes de recherche

Optimisation pour le Machine Learning:

- développement de nouveaux algorithmes “frugaux”
  - moins de données
  - moins de temps de calcul
  - exploitent la structure des données
- science reproductible : benchmarks publics, implémentation open source
- travaux utilisés en neuroscience, géophysique, génomique...

Ce que j’aime dans le ML :

- maths
- code (`https://github.com/mathurinm`), communauté open source
- applis/théorie (neurosciences, musique, image)
- milieu très dynamique

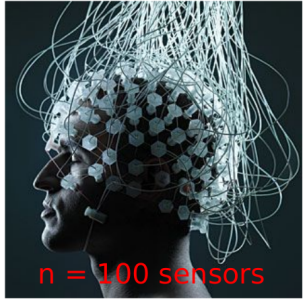
Régularisation

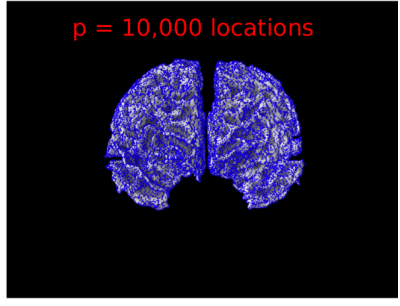
- le vrai but n'est pas de bien prédire sur les données d'entraînement, mais sur des nouvelles entrées
- modèle actuels : suffisamment de paramètres pour apprendre par coeur toutes les données d'entraînement (*surapprentissage*)
- approche classique : coût = erreur + complexité du modèle

on essaye d'apprendre des modèles qui prédisent bien sur les données d'entraînement tout en restant simples

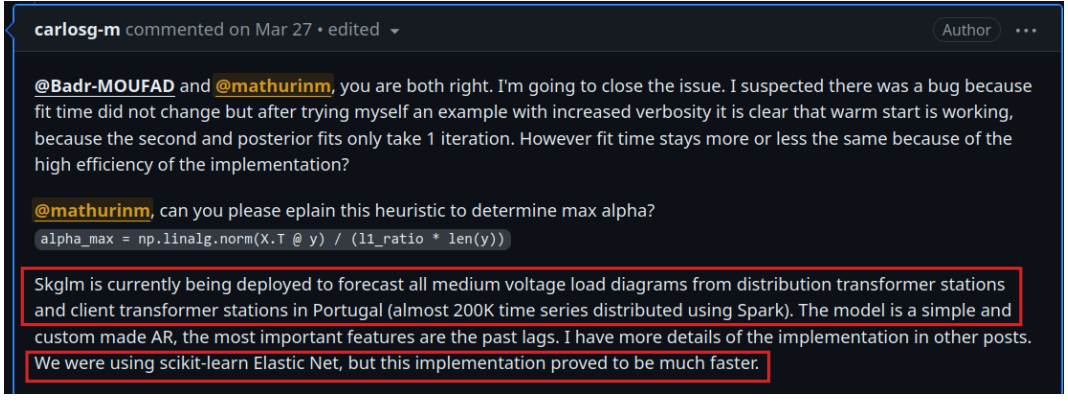
Exemple : la parcimonie

- modèle avec beaucoup de paramètres potentiels
- mais on veut n'en sélectionner qu'une petite fraction
- applications : neurosciences, activité cérébrale





Open source : skglm



https://contrib.scikit-learn.org/skglm/

Open source : skglm

Hi there,

Following up on this after a very long pause, we recently finally got to use skglm and I wanted to say: hats off to you all, it is absolutely brilliant.

It managed to fit an Elastic Net model to a dataset with O(100k) features and 1m rows in 13s(!), with a more than satisfactory precision-recall curve.

For comparison, sklearn's comparable Elastic Net model took 2h to fit only O(100k) rows (we did not have the patience to wait to see how long it might take with 1m rows, it did not complete after many hours) and gave similar performance to skglm.

Even better: tuning the L1/L2 parameters to hopefully be similar, skglm produced a model with O(100) non-zero coefficients, whereas sklearn ended up with a model with O(10k) non-zero coefficients. This may be a mistake on my part in tuning those parameters to be the same, however given the comparable performance and improved parsimony, skglm wins this nevertheless.

Skglm is very likely what we will be switching to in production for this use case.

Travaux actuels

- nouveaux algorithmes pour l'apprentissage de graphes (thèse Can Pouliquen)
- nouvelles architectures de réseaux de neurones (thèse Anne Gagneux)
- modèles génératifs pour l'image

Travaux de l'équipe Ockham :

- transport optimal (Titouan Vayer)
- problèmes inverses (Rémi Gribonval)
- traitement d'images (Marion Foare)
- optimisation frugale (Elisa Riccietti)

Quelques conseils

- contacter les gens (profs, élèves) : vous ne dérangez pas
- la recherche : parfois difficile mais très gratifiant
- sujet $\ll$ encadrement
- co-encadrement : plus de dispo, complémentarité idées/temps

Quelques conseils

- contacter les gens (profs, élèves) : vous ne dérangez pas
- la recherche : parfois difficile mais très gratifiant
- sujet $\ll$ encadrement
- co-encadrement : plus de dispo, complémentarité idées/temps
- écouter tous les conseils et n'en suivre aucun :)