

*Inria*

# Comment faire apprendre à une machine ?

Café gourmand, 08/03/2022

Mathurin Massias (DANTE)

# 01

Qu'est-ce qu'un modèle  
de Machine Learning?

### Modèle de ML : système entrée/sortie

- prend en entrée des données
  - renvoie une sortie (*prédiction*)
- 
- une photo > il y a un chat dessus ?
  - un plateau d'échecs > le meilleur mouvement à effectuer ?
  - un électrocardiogramme:
    - > normal/pathologique ?
    - > liste des pathologies présentes ?
  - un texte en anglais > sa traduction française

variété des types de données, variété des applications

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

- vrais exemples avec la bonne réponse (*supervision*)
- beaucoup
- coûteux
- différence notable avec l'humain

### Pistes d'amélioration actuelles :

- "transfert"
- "autosupervision"

2) On transforme les données en une liste de nombres

- image, signal audio : OK (a priori)
- plateau d'échec : ok
- texte : ?

**Difficulté** : ne pas perdre le sens des données

3) On postule une forme pour la relation entrée/sortie : *choix* du modèle, a priori

Formule mathématique “paramétrique”, par ex:

$$\text{sortie} = a \times \text{entrée nombre1} + b \times \text{entrée nombre2} + \dots$$

Tâche simple: prédire le prix d'une maison à partir de sa surface

On peut postuler un modèle simple :

$$\text{prix}(\text{maison}) = a \times \text{surface}(\text{maison})$$

Modèle à 1 paramètre, en faisant varier  $a$  on fait varier ses prédictions

4) On “optimise” les paramètres du modèle: on trouve leur *bonne* valeur

$$\text{prix} = a \times \text{surface}$$

- ajuste  $a$  par corrections successives
- repose sur un **coût**: comparaison entre la prédiction du modèle et la vraie valeur de sortie

bonne valeur = minimise le coût (l'erreur commise)

Les modèles modernes ont des millions ou des milliards de paramètres :

$$\text{sortie} = a \times \text{entrée1} + b \times \text{entrée2} + \dots$$

et ils font intervenir des fonctions plus compliquées que + et ×  
leur entraînement prend du temps



Jeu de go, traduction, reconnaissance d'image...



### Combinaison de 3 facteurs:

- Données accessibles (capteurs, numérisation...)
- Modèles capables d'apprendre des relations complexes
- Puissance de calcul ++ (matériel + logiciel)

02

Ma recherche

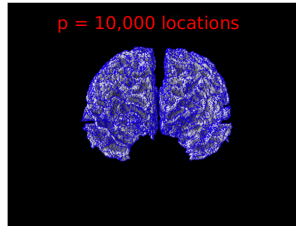
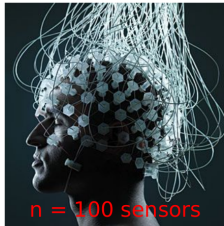
## Optimisation pour le Machine Learning:

- 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...

- 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

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



- comment mieux régulariser ? (moins coûteux, plus facile pour le praticien)
- comment se servir de la régularisation pour résoudre les problèmes d'optimisation plus vite ?
- comment peut-on changer la façon de résoudre le problème d'optimisation pour régulariser ?