# 24 Exemples d'algorithmes d'apprentissage supervisés et non-supervisés

Un ordinateur ne résout pas des problèmes, il <u>calcule</u>. Avant, il faut <u>modéliser</u>. C'est souvent <u>plus dur</u> que de trouver comment calculer.

Exemples de problèmes :

- ▷ Combien de TVA à payer ? ⟶ Modèle (et calcul) facile(s)
- ▷ Quelle sera la météo demain ? ⟶ Modèle physique compliqué
- ▷ Cette photo a-t'elle un panneau STOP ? ⟶ Insondable..?

Principe de l'apprentissage : Des modèles génériques, dépendant de paramètres ; des exemples particuliers pour régler les paramètres.

## I Apprentissage supervisé : exemple de l'algo k-NN

### ① Problème d'apprentissage supervisé : formalisation

On a une collection de N exemples, dont on connait des attributs $X_i \in \mathcal{X}$ l'espace de description, et une propriété $Y_i \in \mathcal{Y}$ l'espace cible. On voudrait construire une fonction $f : \mathcal{X} \to \mathcal{Y}$ qui prédit bien le problème original $f^*$, soit :

pour les instances $\omega$, $\mathbb{E}\left[\mathrm{dist}\left(f(\mathrm{descr}(\omega)), f^*(\omega)\right)\right]$ est minimal

□ Quand $\mathcal{Y}$ est un ensemble fini : problème de classification

### ② k-plus proches voisins (k-NN)

On suppose qu'on possède une distance $\delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$

La fonction ▷ calcule, pour un $X \in \mathcal{X}$, toutes les distances

$$\delta(X, X_j) \text{ pour } j = 1, \dots, N$$

▷ retient les k exemples les plus proches

▷ produit un consensus à partir des $Y_j$ retenus

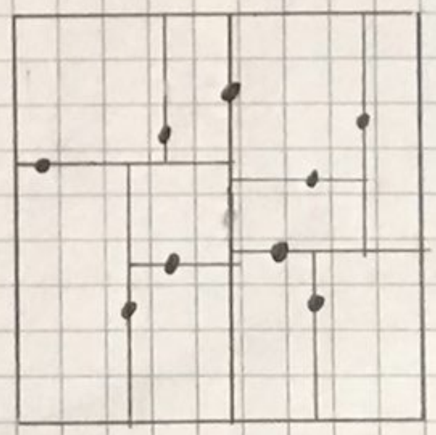
↳ pour une classification, on prend la classe majoritaire

↳ dépend de la pertinence de l'espace $\mathcal{X}$ et de la distance $\delta$.

### ③ Cas euclidien : accélération par K-d tree

Problème : pour chaque appel de $f$, $\Theta(N)$ appels de $\delta$

Si $\mathcal{X} \subset \mathbb{R}^K$ et $m$ est la distance euclidienne $\|\cdot\|_2$, on peut éviter beaucoup de calculs avec un "K-d tree"



Structure : arbre binaire où chaque noeud est la médiane de son sous-arbre, pour une dimension

Principe : on peut minorer la dist à tout un sous-arbre.

↳ $O(\log N)$ calculs pour K fixé (mais malédiction de la dimension !)

□ La distance euclidienne est naturelle (composition des écarts-type)

# II Apprentissage non-supervisé : Clustering, k-moyennes

## ① Clustering : principes, applications

On a seulement des $X_j \in \mathcal{X}$ et une distance $\delta$, on veut les regrouper en un petit nombre de clusters. Objectifs :

- ▷ Segmentation : compresser l'espace $\mathcal{X}$, ou tracer des frontières
- ▷ Profiling : établir des profils-type à traiter différemment (p.e. marketing)
- ▷ Exploration : visualiser, comprendre des données complexes

## ② Problème des k-moyennes

On encode chaque cluster $\mathcal{C}_i$ par un représentant $\overline{X}_i$, et on cherche à minimiser $\Phi = \sum_{i=1}^{k} \Phi_i(\overline{X}_i)$, où $\Phi_i(X)$ l'inertie intra-cluster vaut $\sum_{x \in \mathcal{C}_i} \delta^2(x, X)$.

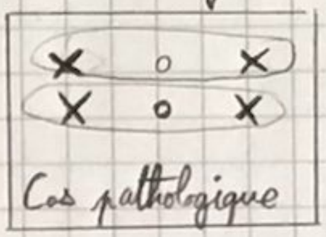
- □ Analogie mécanique : si $\mathcal{X} = \mathbb{R}^3$ et chaque $x \in \mathcal{X}$ a une masse unitaire, alors $\Phi_i(X)$ est l'inertie de rotation de l'ensemble des points de $\mathcal{C}_i$ autour du point $X$ (pour $\delta = \|\cdot\|_2$)
- □ Propriété : dans $\mathbb{R}^d$, $\Phi_i(X) = \|\overline{x_i} - X\|_2^2 + \Phi_i(\overline{x_i})$ où $\overline{x_i}$ est le centre de masse (= la moyenne) de $\mathcal{C}_i$

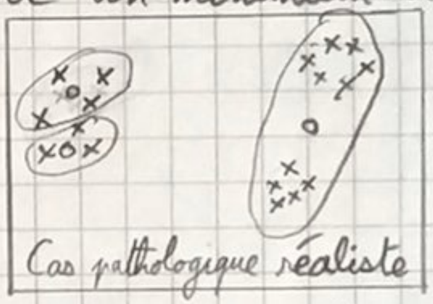
  ↳ à clusters fixés, le $\overline{X_i}$ optimal est $\overline{x_i}$

- □ Le problème de minimiser $\Phi$ est NP-complet

## ③ Résolution de Lloyd

- □ À partition fixée en clusters, on peut optimiser les représentants.
- □ À représentants fixés, on peut optimiser les clusters : un point s'associe au plus proche représentant
- □ Algorithme de Lloyd : alterner ces deux étapes jusqu'à convergence ⇒ Trouve un minimum local.




- □ Problème : un minimum local peut être mauvais.

▷ Développement : Initialisation k-means++, preuve du facteur d'approximation

# III Evaluation et biais

## ① Evaluation, cross-validation

- □ Pour les problèmes supervisés
- □ Problème : on n'a pas d'accès calculatoire à l'espace de base du problème

On estime à partir des échantillons. Mais un échantillon d'évaluation (ie test) ne peut pas être utilisé pour construire le modèle ! (exemple : 1-NN aurait toujours 100% de réussite)

↳ On sépare *a priori* les données d'entraînement et celles de test

- Les modèles ont des meta-paramètres qu'on peut ajuster. (p.e. pour k-NN, le k, et la fonction $\delta$). Les choix d'ajustement dépendent des données d'entraînement, donc il faut une autre couche de protection.

| | | | |
|---|---|---|---|
| train | train | test | validate |
| train | test | train | validate |
| test | train | train | validate |

La cross-validation permet ça sans mobiliser trop de données pour les tests.

## ② Biais statistiques, sur-apprentissage

- À faible ratio échantillon/descripteurs, des "surprises" statistiques
  - ↳ "Selon une étude" porter un t-shirt bleu augmente votre QI. ($n=30$)
  - ▷ Soit limiter la dimension de $X$
  - ▷ Soit contrôler ces effets avec des meta-paramètres
  - ↳ Réseaux de neurones: ces effets ont tendance à apparaître plus tard dans la convergence. Il faut stopper l'apprentissage avant.

## ③ Propriétés des modèles

- Certains algorithmes ne peuvent apprendre que des modèles vérifiant certaines propriétés, ou favoriser ces modèles.
  - ↳ K-means favorise les clusters de taille comparables
  - ↳ Les réseaux de neurone convolutionnels pour l'image sont faits pour être invariants aux décalages, changements de luminosité, rotations, grandissements...
- Ça peut être souhaité, ou non. Mais il faut le savoir.

▷ Développement — Propriétés de k-NN euclidiens et des SVM: exploration avec scikit-learn

## ④ Biais d'échantillonnage, biais sociaux

- Si les données d'entraînement n'échantillonnent pas bien l'espace d'utilisation, le modèle peut mal généraliser
  - ▷ Socrate, Diogène et le poulet
  - ▷ Les photos d'haltère ont souvent un bras qui les tient
  - ↳ Ne se voit pas à la cross-validation et au tests
- Moins drôle: biais discriminatoires (sexistes, racistes, etc.)
  - ↳ p.e. calcul de salaire d'un·e employé·e
  - ▷ Problématique quand utilisé pour une prise de décision
  - ▷ Confusion sur la cible ("statistiques" vs "recommendation")
  - ↳ Écouter les sociologues, contrôler la performance selon des variables de confusion
  - ↳ Travailler l'explicabilité des modèles.