Exemples d'algorithmes d'apprentissage supervisé et non supervisé. Exemples et applications.

I. Algorithmes d'apprentissage

A. Problèmes d'apprentissage [AZE 1.2]

Définition 1 **L'apprentissage** consiste à l'écriture d'algorithme qui renvoie des méthodes de résolutions (grouper des images ensembles) contrairement à l'algorithmique classique qui répond à une requêtes données (trier une liste). On les utilise quand les *données* sont abondantes mais les *connaissances* sont limités.

Définition 2 **Supervisé.** Un problème d'apprentissage est dit supervisé s'il étant donnée n **observations** $x^1, ..., x^n \in \mathcal{X}$ et leurs **étiquettes** $y^1, ..., y^n \in \mathcal{Y}$, on cherche à trouver une fonction $f : \mathcal{X} \to \mathcal{Y}$ tel que $f(x^i) \approx y^i$.

Définition 3 **Classification binaire.** Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est binaire, autrement dit $\mathcal{Y} = \{0, 1\}$ est appelé un problème de **classification binaire**.

Exemple 4 Savoir si un mail est un spam ou non, savoir si un tableau a été peint par Picasso sont des problèmes de classification binaire.

Définition 5 **Classification Multi-classe.** Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est discret et fini, autrement dit $\mathcal{Y} = \{1, 2, ..., C\}$ est appelé un problème de classification multi-classe. C est le nombre de classes.

Exemple 6 **La reconnaissance de chiffres manuscrits** dans une image est un exemple de problème de classification multi-classe d'un ensemble d'images vers $\mathcal{Y} = [0, 9]$.

Exemple 7 **La base MNIST** est une banque de données d'images de chiffres manuscrits annotées par les classes attendues.

Définition 8 **Régression** Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est $\mathcal{Y} = \mathbb{R}$ est appelé un problème de régression.

Définition 9 **L'apprentissage non supervisé** est la branche du machine learning qui s'intéresse aux problèmes pouvant être formalisés de la façon suivante : étant données n observations $\{x\square^i\}_{i=1,...,n}$ décrites dans un espace $\mathcal{X}$, il s'agit d'apprendre une fonction sur $\mathcal{X}$ qui vérifie certaines propriétés.

Définition 10 **Le partitionnement** (ou clustering) est un problème d'apprentissage non supervisé pouvant être formalisé comme la recherche d'une partition $\bigcup_{k=1}^{K} C_k$ des n observations $\{x\square^i\}_{i=1,...,n}$. Cette partition doit être pertinente au vu d'un ou plusieurs critères à préciser.

Exemple 11 Une image $x^i \in \mathcal{X}$ est classiquement représentée comme un vecteur de R^m.

B. Fonctions de coûts [AZE 2.4] et Notions de distances [AZE 8.3]

Définition 12 **Une fonction de coût** $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, aussi appelée fonction de perte ou fonction d'erreur est une fonction utilisée pour quantifier la qualité d'une prédiction. $L(y, f(x))$ est d'autant plus grande que $f(x)$ est éloignée de la vraie valeur y.

Exemple 13 En considérant les étiquettes comme des vecteurs on utilisera classiquement la distance euclidienne.

Définition 14 **Minimisation fonction de coût.** Un algorithme d'apprentissage a pour objectif de minimiser $L(f(x^i), y^i), \forall i$.

Définition 15 **Sur-apprentissage et généralisation.** Simplement minimiser la fonction de coût sur les données connues mènent au phénomène de sur-apprentissage et limite la qualité de la fonction f sur des données non présentes dans les données d'entraînement. On dit que la fonction f n'a pas généraliser.

Définition 16 **Une distance** est une fonction $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, réflexive, nulle entre 2 points égaux et véirfiant l'inégalité triangulaire. Elle permet de comparer les observations sans connaitre leurs étiquettes.

DEV

Exemple 17 Une distance point à point entre deux signaux temporels ne reflètent pas bien l'espace des signaux.

Définition 18 Une méthode de programmation dynamique, dit « Time Warping », permet de définir une fonction de coût plus proche de la topologie réelle des signaux.

II. Algorithme Supervisé [TOR 9.7.1]

A. Algorithmes des k plus proches voisins [AZE 8.2]

Exemple 19 **L'algorithme des plus k plus proches voisins** classifie les nouvelles observations par un mélange des classes de ses k plus proches voisins.

Définition 20 **Choix de classe** en fonctions des k plus proches voisins. On peut choisir la classe majoritaire dans le cas d'un problème de classification, ou la moyenne des classes pour un problème de régression.

Complexité 21 **Algorithme naïf** $\mathcal{O}(n)$ pour calculer les n distances et $\mathcal{O}(n.\log(k))$ pour trouver les k plus petites.

Définition 22 **Un arbre k-dimensionnel** est un arbre binaire de recherche partitionnant successivement une dimension de l'espace par profondeur.

Complexité 23 **L'utilisation d'un arbre k-dimensionnel** permet de réduire la complexité d'une prédiction des k plus proches voisins à $O\left(kn^{1-\frac{1}{k}}\right)$ en moyenne (admis).

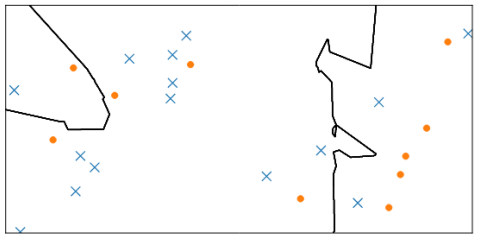
Définition 24 **Une matrice de confusion** est une matrice où $M_{i,j}$ est le nombre de données classées comme j par l'algorithme et dont la classe réelle est i. On peut mesurer le taux d'erreur comme la proportion de données hors de la diagonale.

Définition 25 **Hyperparamètre k** [AZE 8.2.3]

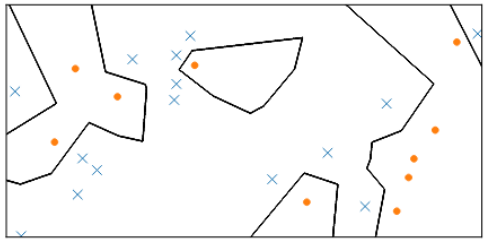
- si $k = 1$, l'algorithme très sensible au bruit, risque de sur-apprentissage.
- si $k = n$, on prédit toujours la classe majoritaire de notre jeu de données.

On utilise en général une valeur intermédiaire. L'heuristique $k = \sqrt{n}$ est parfois utilisée.

Exemple 26 **Classification 5 plus proches voisins** [AZE 8.3] .



Exemple 27 **Exemple de sur-apprentissage avec $k = 1$.**



B. Apprentissage hiérarchique [AZE]

Définition 28 **Un modèle hiérarchique** se comporte comme une suite de tests conditionnels.

Définition 29 **Un arbre de décision** est un modèle hiérarchique pouvant être représenté sous forme d'un arbre. Chaque nœud corresponds à une décision (oui/non). Les feuilles correspondent à une étiquette.

Exemple 30 **Le jeu du « Qui-est-ce? »** peut-être représenté par un arbre de décision. Chaque question posée permet de raffiner un sous-ensemble de données satisfaisant une suite de critères.

Définition 31 **L'entropie de Shannon** est une mesure de la quantité d'incertitude mesurée en bits sur un ensemble de données S. Elle est donnée par la formule $H(S) = \sum_{x \in S} -p(x)\log_2 p(x)$. Plus l'entropie de S est grande, plus les données présentes sont variées.

DEV

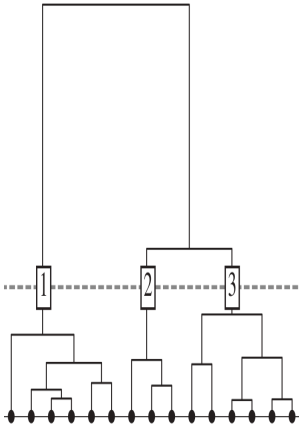
Algorithme 32 **ID3** [TOR 9.7.1] est un algorithme permettant la construction d'un arbre de décision en choisissant successivement le critère le plus discriminant au sens de l'entropie (i.e. ayant la plus faible entropie) sur un ensemble de données.

III. Algorithme Non Supervisé [TOR 9.7.2]

A. Clustering Ascendant Hiérarchique [AZE 12.3]

Définition 33 **La distance entre cluster** peut être défini comme le minimum de distance entre tous les éléments de deux clusters.

Algorithme 34 **La méthode ascendante** consiste à chaque étape à fusionner les deux clusters à distance minimal. On démarre avec un cluster par observation.



Définition 35 Le résultat d'un clustering hiérarchique peut se visualiser sous la forme d'un **dendrogramme**. Il s'agit d'un arbre binaire dont les n feuilles correspondent chacune à une observation. Chaque nœud correspond à un cluster.

Remarque 36 Avantage de ne pas avoir à choisir le nombre de cluster du clustering car la dendogramme les stockent tous.

Complexité 37 $\mathcal{O}(n^3)$, n étapes de calcul des n^2 pairs de distance.

B. Algorithme des k moyennes [AZE 12.4]

Définition 38 On appelle centroïde du cluster C le point défini par : $\mu_C = \frac{1}{|C|} . \sum_{x \in C} x$.

DEV

Algorithme 39 de Loyd.

1. Choisir des centroïdes initiaux parmi les observations.
2. Affecter chaque observation au centroïde dont elle est le plus proche.
3. Recalculer les centroïdes de chaque cluster.
4. Répéter les opérations 2-3 jusqu'à convergence, i.e. jusqu'à que les affections ne changent plus.

Remarque 40 **Convergence (Admis).**

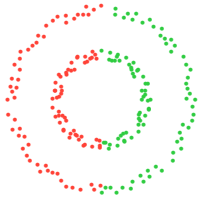
Remarque 41 Les clusters formés par l'algorithme des K-moyennes forment un diagramme de Voronoï et sont toutes convexes.

Remarque 42 **Les points aberrants** vont généralement « attirer » le cluster vers eux seuls et doivent généralement être pris en compte séparément.

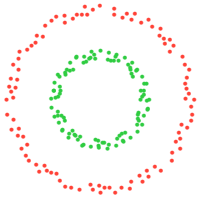
Remarque 43 Une variante nommé k-moyenne++ de cette algorithme consiste à choisir de façon déterministe les centroïdes initiaux pour les séparer un maximum.

Exemple 44

k-moyenne
$k = 2$



clustering
ascendant
hiérarchique



IV. Enjeux éthiques et sociétaux

A. Biais des algorithmes d'apprentissages

Enjeux 45 **Des biais dans les données** formeront également des biais similaire dans la classification renvoyée par un algorithme d'apprentissage. Il est nécessaires de s'assurer d'une représentation équilibrée des différentes classes attendues dans les données d'entraînement pour éviter ce phénomène.

Définition 46 **Un modèle est explicable** s'il est capable de fournir des justifications claires sur la classification renvoyée.

Définition 47 **Un modèle est interprétable** si l'on est capable d'en comprendre le fonctionnement et les raisons pour lesquelles il classe les données d'une certaine manière.

B. Utilisation des données personnelles

Définition 48 **Le RGPD** (protection des données) est une loi européenne écrite en 2016 qui encadre la gestion des données personnelles. Les entreprises doivent justifier de l'utilisation de la récolte des informations personnelles et les garder un temps limité.

Exemples d'algorithmes d'apprentissage supervisé et non supervisé. Exemples et applications.

I. Algorithmes d'apprentissage

A. Problèmes d'apprentissage [AZE 1.2]

1 Def L'apprentissage
2 Def Supervisé.
3 Def Classification binaire.
4 Ex
5 Def Classification Multi-classe.
6 Ex La reconnaissance de chiffres manuscrits
7 Ex La base MNIST
8 Def Régression
9 Def L'apprentissage non supervisé
10 Def Le partitionnement
11 Ex

B. Fonctions de coûts [AZE 2.4] et Notions de distances [AZE 8.3]

12 Def Une fonction de coût
13 Ex
14 Def Minimisation fonction de coût.
15 Def Sur-apprentissage et généralisation.
16 Def Une distance
17 Ex
18 Def

II. Algorithme Supervisé [TOR 9.7.1]

A. Algorithmes des k plus proches voisins [AZE 8.2]

19 Ex L'algorithme des plus k plus proches voisins
20 Def Choix de classe
21 Complex Algorithme naïf
22 Def Un arbre k-dimensionnel
23 Complex L'utilisation d'un arbre k-dimensionnel
24 Def Une matrice de confusion
25 Def Hyperparamètre k [AZE 8.2.3]
26 Ex Classification 5 plus proches voisins [AZE 8.3].
27 Ex Exemple de sur-apprentissage avec $k = 1$.

B. Apprentissage hiérarchique [AZE]

28 Def Un modèle hiérarchique
29 Def Un arbre de décision
30 Ex Le jeu du « Qui-est-ce? »
31 Def L'entropie de Shannon
32 Algo ID3 [TOR 9.7.1]

III. Algorithme Non Supervisé [TOR 9.7.2]

A. Clustering Ascendant Hiérarchique [AZE 12.3]

33 Def La distance entre cluster
34 Algo La méthode ascendante
35 Def Dendogramme.
36 Rem
37 Complex

B. Algorithme des k moyennes [AZE 12.4]

38 Def
39 Algo de Loyd.
40 Rem Convergence (Admis).
41 Rem
42 Rem Les points aberrants
43 Rem
44 Ex

IV. Enjeux éthiques et sociétaux

A. Biais des algorithmes d'apprentissages

45 Enjeux Des biais dans les données
46 Def Un modèle est explicable
47 Def Un modèle est interprétable

B. Utilisation des données personnelles

48 Def Le RGPD

Remarque

- ne pas parler de réseaux de neurones / deep learning car hors programme
- 4 algorithmes au programme il faut bien les avoirs en tête (chacun peut être un dev) : ID3, k d tree, k moyenne et arbre ascendant hierarchique
- penser à dire que minimiser une fonction de coût peut signifier différentes choses. On peut minimiser la somme des coûts, le minimum des coûts ...
- **Transition III.A → III.B** : Faire comprendre que l'on souhaite obtenir un algorithme plus rapide mais pour lesquels on va fixer le nombre de cluster (pas toujours facile de choisir cet hyperparamètre).
- Bien clarifier le fait que l'utilisation du « k » est ambigu mais c'est l'usage pour ces algorithmes
- Bien différencier « distance » (pour les observations, non supervisé) et « fonction de coût » (pour les étiquettes, supervisé).
- **Supposition** : le calcul de distance est en $\mathcal{O}(1)$. Si ce n'est pas le cas on multiplie simplement toutes les complexités par ce facteurs.
- Exemple clustering ascendant hierarchique : distance entre cluster à définir précisément à l'oral (minimum des distances entre tous les points).

Bibliographie

[AZE] C. Azencott, *Introduction au Machine Learning.*

[TOR] T. Balabonski & S. Conchon & J. Filliâtre & K. Nguyen & L. Sartre, *MP2I MPI, Informatique Cours et exercices corrigés.*

Paul ADAM & Aloïs Rautureau