Analyse de données

Classification

Rakotoarimalala Tsinjo Tony

ITU 2023

Contexte

- On veut enseigner à un système à savoir reconnaître un chiffre dans une image.
- Les images sont les individus
- Les chiffres sont les classes. Ici les classes sont au nombre de 10 : $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
- Le système, après avoir appris, pourra donc prendre en entrée une image et donner en sortie la classe correspondante

Entrée

$\Rightarrow$ Sortie

1

Apprentissage

- On utilise alors un ensemble de données d'entraînement
- Cet ensemble est un ensemble de couple (x_i, y_i) où x_i représente un individu et y_i représente la classe de x_i.
- A partir de ces données, selon le modèle choisi, le système va mettre en place une représentation de la connaissance à donner une classe/étiquette à un x.
- Cette représentation peut être
  1. sous forme de graphe ou arbre
  2. sous forme d'une fonction bien définie (une droite/hyperplan, polynôme...)
  3. sous forme d'ensemble de poids d'un réseau
- Pour notre exemple, les données d'entraînement sont des couples (x, y) où x représente une image et y la classe, pour nous c'est le nombre représente par l'image

Inférence

- la partie inférence consiste à donner à notre modèle une nouvelle donnée qui n'est pas présent dans les données d'entraînement
- Le système en utilisant la connaissance apprise dans la phase d'entraînement pour prédire la classe correspondante à cette nouvelle donnée.
- Pour notre exemple, on va donner une nouvelle image d'un chiffre et le système va prédire le nombre écrit sur l'image

Evaluation du modèle

- On va prendre des nouvelles données pour lesquelles on connaît les étiquettes/classes
- On donne au modèle/système qui a appris, ces données sauf qu'on ne lui fournit pas les étiquettes
- Il va prédire les étiquettes de ces données
- On va alors comparer les étiquettes prédites et les étiquettes réelles pour avoir la précision de la connaissance acquise

Définition

- La classification est une tâche qui nécessite l'utilisation d'algorithmes d'apprentissage automatique qui apprennent à attribuer une étiquette de classe aux exemples du domaine problématique.
- Un exemple facile à comprendre est la classification des courriels comme "*spam*" ou "*non spam*".

Formalisation

- Soit $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ ensemble de données tel que x_i appartient à un espace $\mathcal{X}$ (par exemple $\mathcal{X} = \mathbb{R}^d$) et $y_i \in \mathcal{Y}$ l'étiquette associée à x_i ($\mathcal{Y}$ étant un ensemble fini).

Catégories de classification

Les différents types de problème de classification dépendent de la taille de $\mathcal{Y}$:

- **classification binaire** si $|\mathcal{Y}| = 2$ (ex: $\mathcal{Y} = \{-1, 1\}$ ou $\mathcal{Y} = \{0, 1\}$).
  Applications : Détection de fraudes, détection d'anomalies, Spam ou pas, ...
- **classification multi-classe** si $\mathcal{Y} = \{1, 2, \ldots K$
  Applications :reconnaissance d'objets, d'écriture
- **classification multi-étiquettes** si $\mathcal{Y} = 2^{\{1,2,\ldots K\}}$
  Applications: Reconnaissance du topics de documents

Comment

On cherche à apprendre une fonction de classification $f : \mathcal{X} \to \mathcal{Y}$ permettant de prédire le label de x.

Classification binaire

La classification binaire fait référence aux tâches de classification en deux étiquettes de classe.
En voici quelques exemples :

- Détection de spam par email (spam ou non).
- Prédiction de désabonnement (désabonnement ou non).
- Prédiction de la conversion (achat ou non).

Classification binaire

La classification binaire fait référence aux tâches de classification en deux étiquettes de classe.
En voici quelques exemples :

- Détection de spam par email (spam ou non).
- Prédiction de désabonnement (désabonnement ou non).
- Prédiction de la conversion (achat ou non).

En général, les tâches de classification binaire impliquent une classe qui représente l'état normal et une autre classe qui représente l'état anormal.

Exemple

Jour	Attributs des exemples				Classe
	Prévisions	Température	Humidité	Vent	
1	Ensoleillé	Chaud	Élevée	Faible	Non
2	Ensoleillé	Chaud	Élevée	Fort	Non
3	Nuageux	Chaud	Élevée	Faible	Oui
4	Pluvieux	Moyen	Élevée	Faible	Oui
5	Pluvieux	Frais	Normale	Faible	Oui
6	Pluvieux	Frais	Normale	Fort	Non
7	Nuageux	Frais	Normale	Fort	Oui
8	Ensoleillé	Moyen	Élevée	Faible	Non
9	Ensoleillé	Frais	Normale	Faible	Oui
10	Pluvieux	Moyen	Normale	Faible	Oui
11	Ensoleillé	Moyen	Normale	Fort	Oui
12	Nuageux	Moyen	Élevée	Fort	Oui
13	Nuageux	Chaud	Normale	Faible	Oui
14	Pluvieux	Moyen	Élevée	Fort	Non

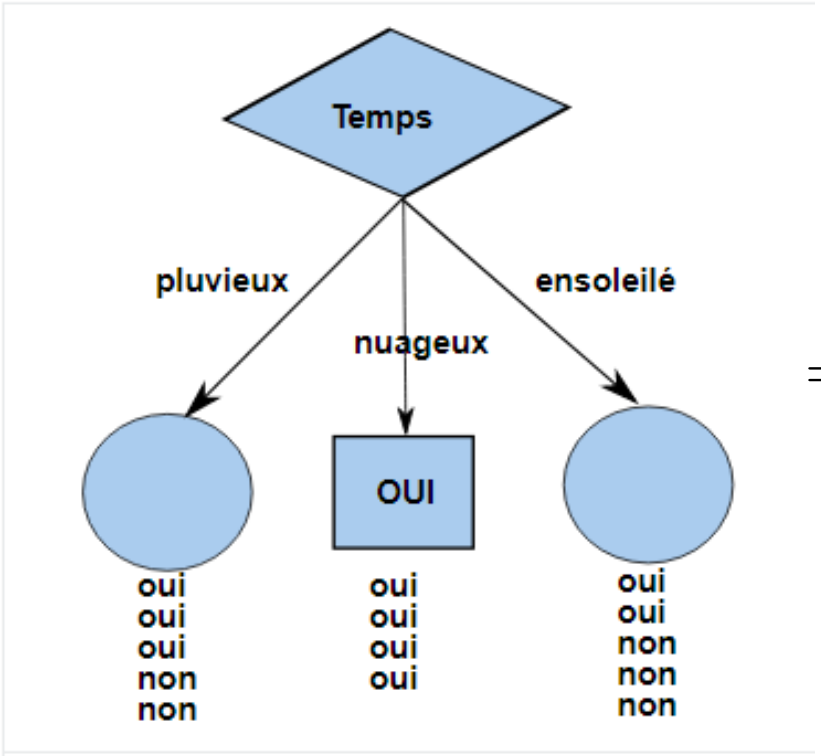
Ensemble d'exemples pour playTennis

Exemple

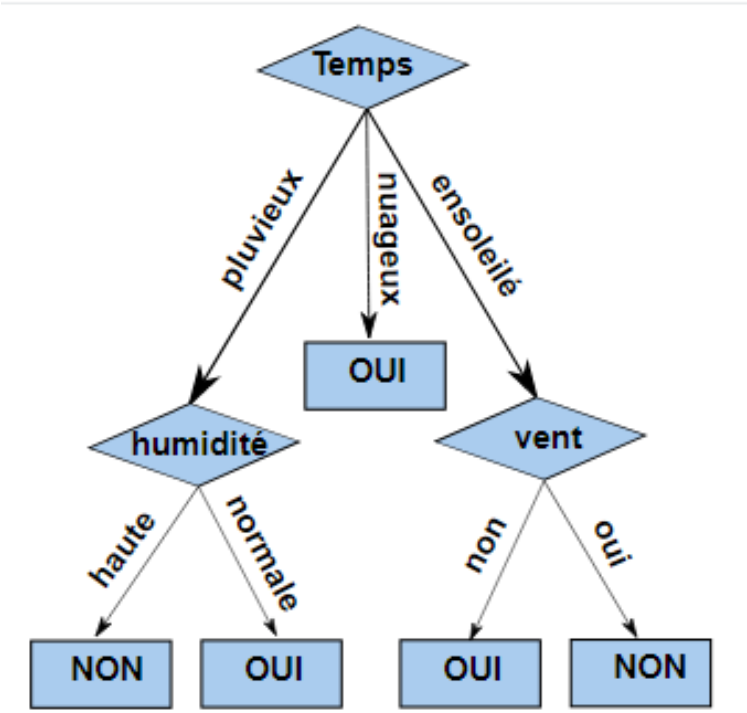
- On veut représenter donc la connaissance quand-est-ce qu'on peut jouer?
- On a une classification binaire avec les classes possibles sont **OUI** et **NON**
- Les individus sont des vecteurs de 4 valeurs

Arbre de décision

- Étant donnée plusieurs caractéristiques (variable), la décision se commence par un de ces caractéristiques; si ce n'ai pas suffisant, on utilise une autre, ainsi de suite.
- Exemple de construction d'un arbre de décision pour notre exemple



$\Rightarrow$



Construction d'un arbre de décision

L'algorithme général de création d'un arbre de décision:

1. Déterminer la meilleure caractéristique dans l'ensemble de données d'entrainement.
2. Diviser les données d'entrainement en sous-ensembles contenant les valeurs possibles de la meilleure caractéristique.
3. Générez de manière récursive de nouveaux arbres de décision en utilisant les sous-ensembles de données créés.
4. Lorsqu'on ne peut plus classifier les données, on s'arrête.

Algo ID3

- ID3 est un algorithme de création d'un arbre de décision.
- L'algorithme général définit ci-dessus reste vrai.
- Il nous reste à expliquer la façon comment ID3 fait pour choisir la meilleure caractéristique/variable pour l'utiliser pour créer un nœud dans l'arbre selon cette variable

Définition

- On appelle entropie d'un ensemble d'individu S

$$\text{Entropie}(S) = \sum_{c \in \text{classes}(S)} -p_c \times \log_2(p_c)$$

- Pour notre exemple, au début S étant l'ensemble des 14 individus donc

$$\text{Entropie}(S) = -\left(\frac{9}{14}\right) \times \log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \times \log_2\left(\frac{5}{14}\right) = 0.94$$

- S'il n'y qu'une seule classe dans S alors

$$\text{Entropie}(S) = 0$$

Définition

- On définit le gain d'entropie d'une variable/caractéristique A sur l'ensemble des individus S :

$$\text{Gain}(S, A) = \text{Entropie}(S) - \sum_{v \in \text{valeurs}(A)} \frac{|S_v|}{|S|} \times \text{Entropie}(S_v)$$

- Pour notre exemple, si on prend $A =$ "*Prévisions*"

temps	jouer(oui)	jouer(non)	S_v	Entropie(S_v)
ensoleilé	2	3	5/11	$\left(-\frac{3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right)\right)$
nuageux	4	0	4/11	$\left(-\frac{0}{4} \times \log_2\left(\frac{0}{4}\right) - \frac{4}{4} \times \log_2\left(\frac{4}{4}\right)\right)$
pluvieux	3	2	5/11	$\left(-\frac{3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right)\right)$

Donc

$$\text{Gain}(S, A) = 0.94 - 0.357 \times (0.97) - 0.286 \times (0) - 0.357 \times (0.97) = 0.24742$$

Choix de la meilleure variable

- On fait les mêmes calculs pour chaque varible
- On obtient alors

	Prévisions	température	humidité	vent
Gain	0.247	0.029	0.152	0.048

- On choisit alors **Prévisions** comme variables pour le premier nœud
- On refait comme le l'algorithme général l'indique