Intelligence Artificielle

Les arbres de décisions

Maria Malek

Département Informatiques

Extraire les connaissances des données

- La classification : chaque donnée est affectée d'une caractéristique.

Extraire les connaissances des données

- La classification : chaque donnée est affectée d'une caractéristique.
- La segmentation : On dispose d'un ensemble de points, la tâche consiste à repérer des groupes de points qui se ressemblent.

Extraire les connaissances des données

- La classification : chaque donnée est affectée d'une caractéristique.
- La segmentation : On dispose d'un ensemble de points, la tâche consiste à repérer des groupes de points qui se ressemblent.
- Recherche des association : On recherche de régularité ou de co-occurrence entre certaines valeurs d'attributs.

La problématique de la classification

- **Une donnée** est un enregistrement,un individu (statistique), une instance (orienté objet), un point, un vecteur.

La problématique de la classification

- **Une donnée** est un enregistrement,un individu (statistique), une instance (orienté objet), un point, un vecteur.
- **Un attribut** peut être de nature qualitative ou quantitative, ou même un enregistrement (comme la date).

La problématique de la classification

- **Une donnée** est un enregistrement,un individu (statistique), une instance (orienté objet), un point, un vecteur.
- **Un attribut** peut être de nature qualitative ou quantitative, ou même un enregistrement (comme la date).
- La classification :
  - Construction d'un modèle arborescent permettant de prédire la classe d'une donnée.
  - Estimation directe de la classe d'une donnée en fonction des exemples.
  - Construction d'un modèle interprétable par un humain (les réseaux de neurones, et les machines à vecteurs supports).

Validation d'un classeur

- L'erreur d'un classeur Er est la probabilité que ce classeur ne prédise pas correctement la classe d'une donnée.

Validation d'un classeur

- L'erreur d'un classeur Er est la probabilité que ce classeur ne prédise pas correctement la classe d'une donnée.
- Le taux de succès est $1 - Er$.

Validation d'un classeur

- L'erreur d'un classeur Er est la probabilité que ce classeur ne prédise pas correctement la classe d'une donnée.
- Le taux de succès est $1 - Er$.
- L'erreur apparente $Erapp$ est mesurée sur les exemples utilisés pour la construction du classeur.

Validation d'un classeur

- L'erreur d'un classeur Er est la probabilité que ce classeur ne prédise pas correctement la classe d'une donnée.
- Le taux de succès est $1 - Er$.
- L'erreur apparente $Erapp$ est mesurée sur les exemples utilisés pour la construction du classeur.
- Estimer la qualité d'un classeur :
  - L'ensemble d'apprentissage X_{app} ;
  - L'ensemble de test X_{test} qui permet d'estimer l'erreur de classification ; on connaît la classe de chaque exemple dans cet l'ensemble.

Les approches inductives

- Méthode d’apprentissage à partir d’exemples : arbres de décision,

Les approches inductives

- Méthode d’apprentissage à partir d’exemples : arbres de décision,
- une feuille indique un classe,

Les approches inductives

- Méthode d’apprentissage à partir d’exemples : arbres de décision,
- une feuille indique un classe,
- un nœud spécifie un test que doit subir un certain attribut, chaque branche sortant de ce nœud correspond à une valeur possible de l’attribut en question.

Les approches inductives

- Méthode d’apprentissage à partir d’exemples : arbres de décision,
- une feuille indique un classe,
- un nœud spécifie un test que doit subir un certain attribut, chaque branche sortant de ce nœud correspond à une valeur possible de l’attribut en question.
- un ensemble qui couvre le domaine et inclut les cas rares,

Les approches inductives

- Méthode d’apprentissage à partir d’exemples : arbres de décision,
- une feuille indique un classe,
- un nœud spécifie un test que doit subir un certain attribut, chaque branche sortant de ce nœud correspond à une valeur possible de l’attribut en question.
- un ensemble qui couvre le domaine et inclut les cas rares,
- un ensemble qui ne contient pas de redondance.

Les arbres de décisions : exemple - 1

NUM	CIEL	TEMP.	HUMI.	VENT	CLASSE
1	ensoleillé	élevé	forte	non	N
2	ensoleillé	élevé	forte	oui	N
3	couvert	élevé	forte	non	P
4	pluvieux	moyenne	forte	non	P
5	pluvieux	basse	normale	non	P
6	pluvieux	basse	normale	oui	N
7	couvert	basse	normale	oui	P

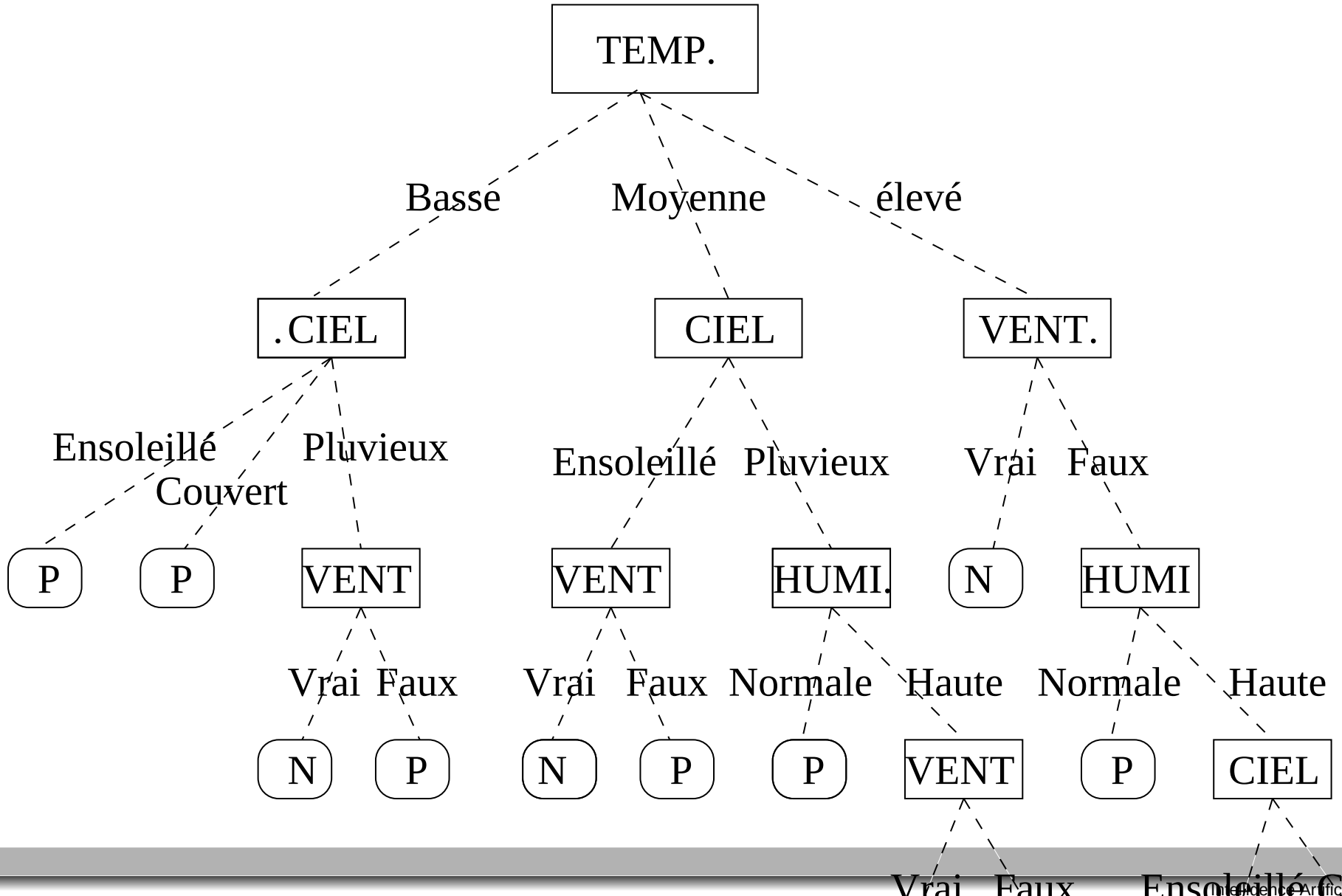
Table 1: Description des conditions météorologiques

Les approches inductives : exemple - 2

NUM	CIEL	TEMP.	HUMI.	VENT	CLASSE
8	ensoleillé	moyenne	forte	non	N
9	ensoleillé	basse	normale	non	P
10	pluvieux	moyenne	normale	non	P
11	ensoleillé	moyenne	normale	oui	P
12	couvert	moyenne	forte	oui	P
13	couvert	élevé	normale	non	P
14	pluvieux	moyenne	forte	oui	N

Table 2: Description des conditions météorologiques

Un arbre de décision !?



Application de l'algorithme ID3

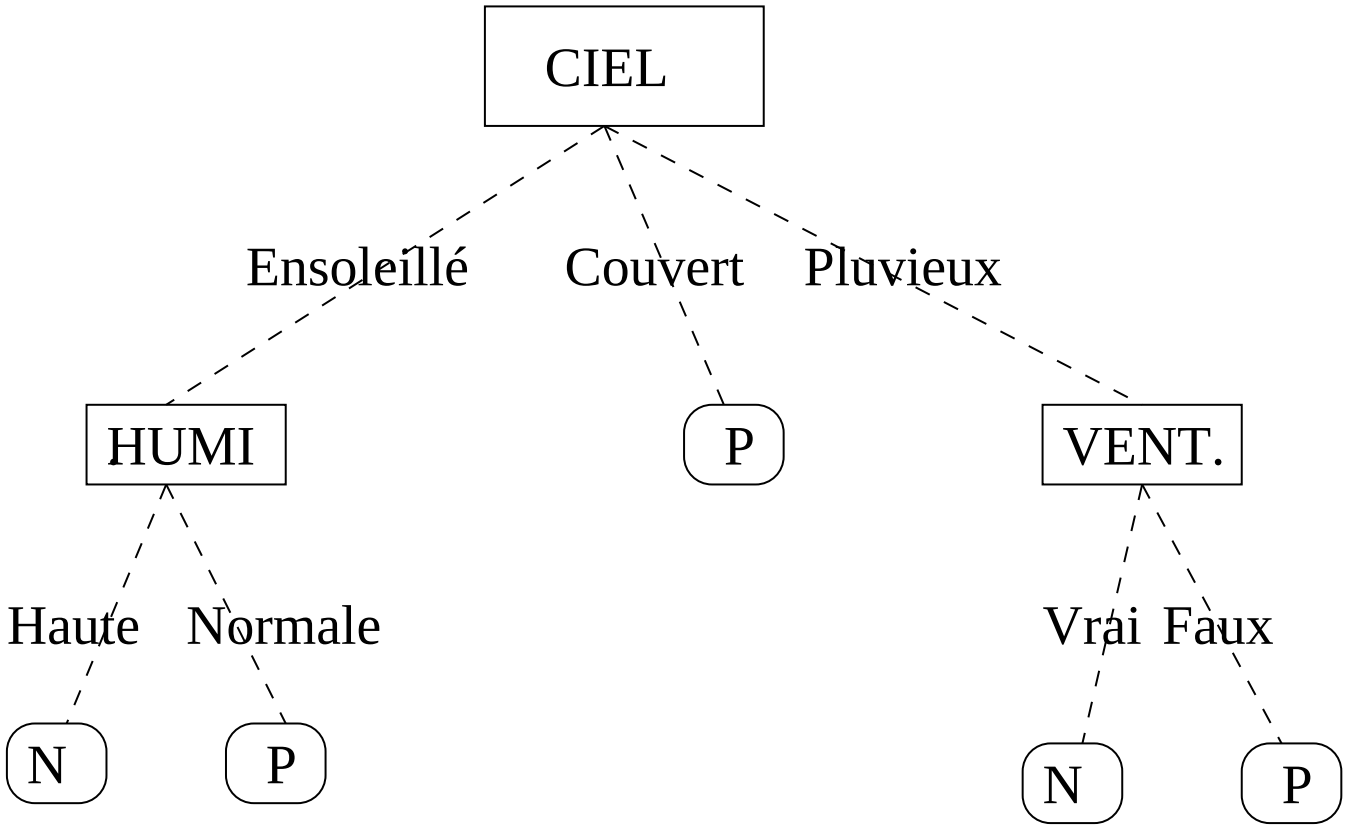


Figure 2: Application de l'algorithme ID3 sur l'ensemble d'apprentissage

Interprétation de l'arbre de décision

- Un arbre de décision exprime un ensemble de règles propositionnelles :
  - SI CIEL= Ensoleillé & HUMI=Forte ALORS CLASSE=N
  - SI CIEL= Ensoleillé & HUMI=Normale ALORS CLASSE=P
  - SI CIEL= Couvert ALORS CLASSE=P
  - SI CIEL= Pluvieux & VENT=Oui ALORS CLASSE=N
  - SI CIEL= Pluvieux & VENT=Non ALORS CLASSE=P

Interprétation de l'arbre de décision

- Un arbre de décision exprime un ensemble de règles propositionnelles :
  - SI CIEL= Ensoleillé & HUMI=Forte ALORS CLASSE=N
  - SI CIEL= Ensoleillé & HUMI=Normale ALORS CLASSE=P
  - SI CIEL= Couvert ALORS CLASSE=P
  - SI CIEL= Pluvieux & VENT=Oui ALORS CLASSE=N
  - SI CIEL= Pluvieux & VENT=Non ALORS CLASSE=P
- le concept Classe=P est décrit par :
  - $(CIEL = Ensoleille \wedge HUMI = Normale) \vee (CIEL = Couvert) \vee (CIEL = Pluvieux \wedge VENT = Non)$

L'algorithme ID3 -1

- Étant donné un ensemble d'apprentissage $E = \{e_i\}$:

L'algorithme ID3 -1

- Étant donné un ensemble d'apprentissage $E = \{e_i\}$:
- chaque exemple est défini par un vecteur d'attributs, et sa classe k_i.

L'algorithme ID3 -1

- Étant donné un ensemble d'apprentissage $E = \{e_i\}$:
- chaque exemple est défini par un vecteur d'attributs, et sa classe k_i.
- A chaque attribut on associe un test T_i :

$$T_i : E \rightarrow V_j = \{V_{ji}\}$$

L'algorithme ID3 -1

- Étant donné un ensemble d'apprentissage $E = \{e_i\}$:
- chaque exemple est défini par un vecteur d'attributs, et sa classe k_i.
- A chaque attribut on associe un test T_i :

$$T_i : E \rightarrow V_j = \{V_{ji}\}$$

- Exemple : test sur l'attribut ciel dans l'ensemble E : donné par : $T_1 : E \rightarrow \{ensoleille, couvert, pluvieux\}$.

L'algorithme ID3 -2

- FONCTION construire-arbre(E:exemples,{T1,..,Tn}:tests): arbre de décision
  - SI *les exemples sont tous d'une même classe* $construire - arbre \leftarrow$ la feuille étiquetée par cette classe
  - SINON SI $T_i \leftarrow bon - choix(\{T_1, .., T_n\})$
    - soit $\{E_{i1}, .., E_{im}\}$ la partition de E pour T_i
    - $construire - arbre \leftarrow$ arbre de racine T_i
    - $l \leftarrow 1$
    - TANTQUE $l <= m$
      - $sous - arbres_l \leftarrow construire - arbre(E_{il}, \{T_1, .., T_n\} \backslash T_i)$
      - $l \leftarrow l + 1$

L'algorithme ID3 -3

- Bon choix ??
  - La quantité d'information donnée par un sous-ensemble de cardinalité n d'un ensemble de N événements est :

$$Q = \log_2 \frac{N}{n} \tag{1}$$

  - L'entropie d'une partition E_1,..,E_k d'un ensemble E de cardinalité N avec $card(E_i) = n_i$ est :
    $I(partition) = \sum_{i=1}^{k} (\frac{n_i}{N}) \log_2 \frac{N}{n_i}$
  - $p_i = \frac{n_i}{N}$:
    $I(partition) = \sum_{i=1}^{k} (p_i) \log_2 \frac{1}{p_i}$

L'algorithme ID3 -4

- Problème de classification avec deux classe, un arbre de décision est un source d'information qui émet :

$$I(n,p) = (\frac{p}{p+n})\log_2(\frac{p+n}{p}) + (\frac{n}{p+n})\log_2(\frac{p+n}{n}) \quad (2)$$

- L'information moyenne fournie par les sous-arbres d'un arbre de racine T est :

$$E(T) = \sum_{i=1}^{k} (\frac{p_i + n_i}{p+n}) I(p_i, n_i) \quad (3)$$

- Le gain d'information pour un test T est défini par : $gain(T) = I(p,n) - E(T)$

L’algorithme ID3 -5

- Exemple (JouerTennis)
  - gain(CIEL) = 0.246 bits,
  - gain(TEMP.) = 0.029 bits,
  - gain(HUMI.) = 0.151 bits,
  - gain(VENT) = 0.048 bits

L'algorithme C4.5 - 1

- Descendant de ID3 ..
  - Normalisation du gain :

$$SplitInfo(T) = -\sum_{i=1}^{n} \frac{\|T_i\|}{\|T\|} \log_2 \frac{\|T_i\|}{\|T\|} \quad (4)$$

le gain est normalisé de la façon suivante :

$GainRatio(T) = \frac{gain(T)}{SplitInfo(T)}$

  - Traitement des valeurs continues : des méthodes pour la discrétisation des valeurs continues ont été ajoutées.
  - Groupement des valeurs des attributs :

L'algorithme C4.5 - 2

- Traitement des valeurs manquantes : supposons que la valeur de l'attribut a_i est manquante :
  - l'ensemble des exemples est organisé en n groupes selon les n valeurs possibles de a_i ;
  - l'exemple est affecté à chacun de ces groupes avec une probabilité d'appartenance selon la fréquence de l'occurrence de cette valeur.
  - cette probabilité est propagée dans l'arbre, elle est en plus prise en compte lors du calcul du gain.
- Élagage de l'arbre de décision : se débarrasser de quelques branches des sous-arbres.