

Apprentissage automatique



Arbres de décision

Intérêts des arbres de décision

☛ Expressivité

- approximation de fonctions à valeurs discrètes
- capable d'apprendre des expressions disjonctives

☛ Lisibilité

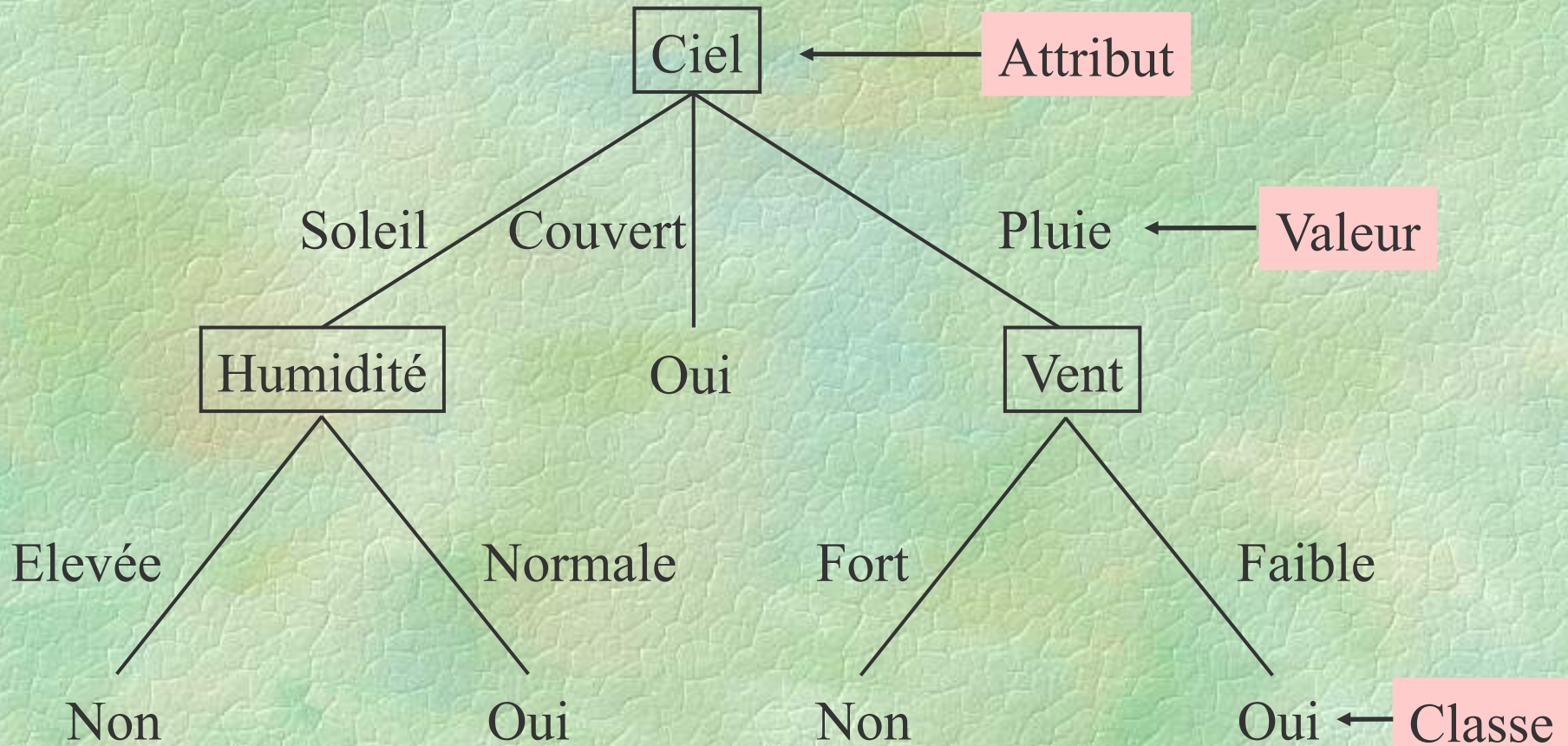
- peut être traduit sous la forme de règles

☛ Beaucoup d'applications

Plan

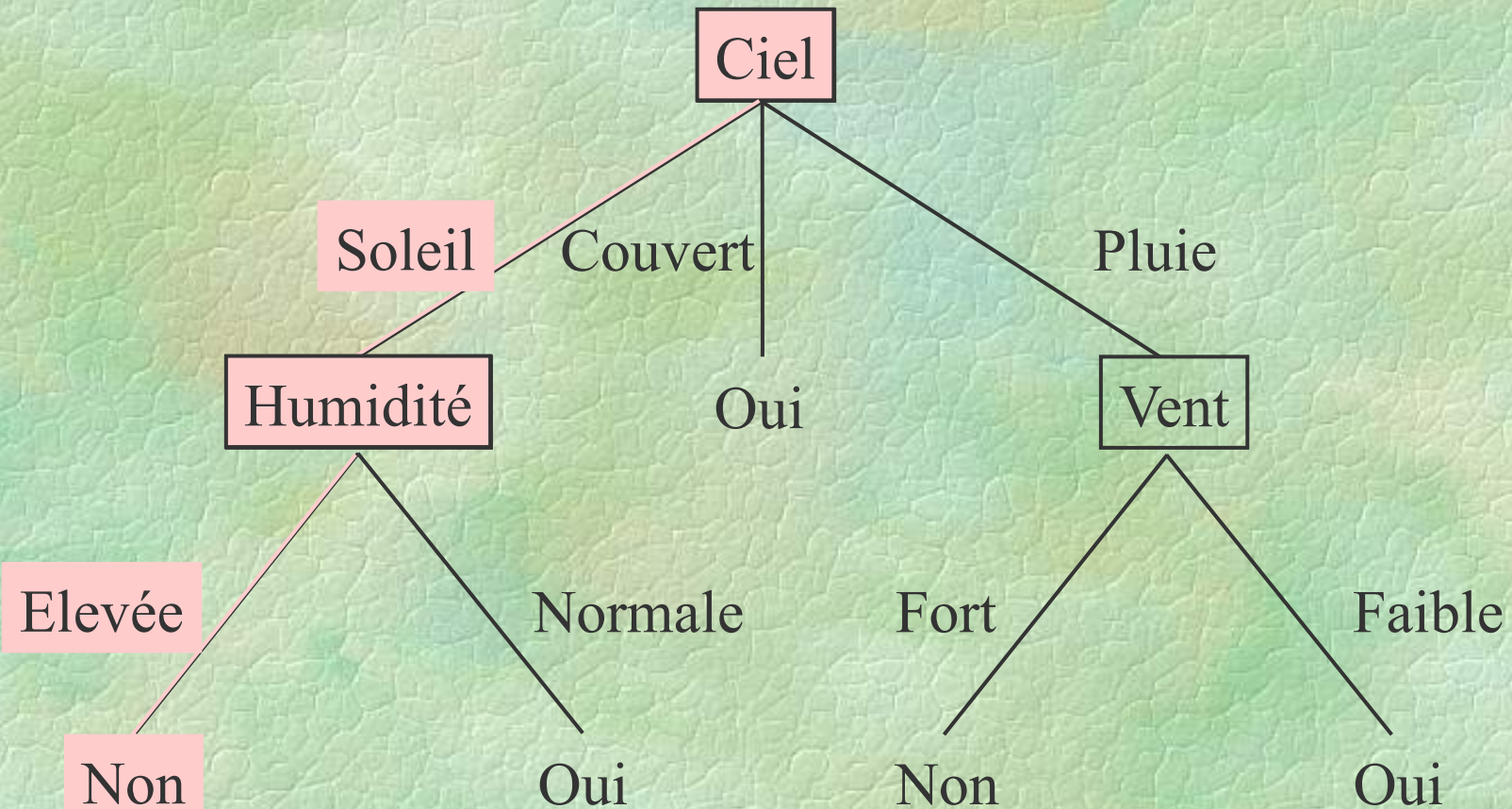
- ☛ Représentation
- ☛ ID3
- ☛ Espace des hypothèses et biais inductif
- ☛ Extensions

Exemple d'arbre de décision

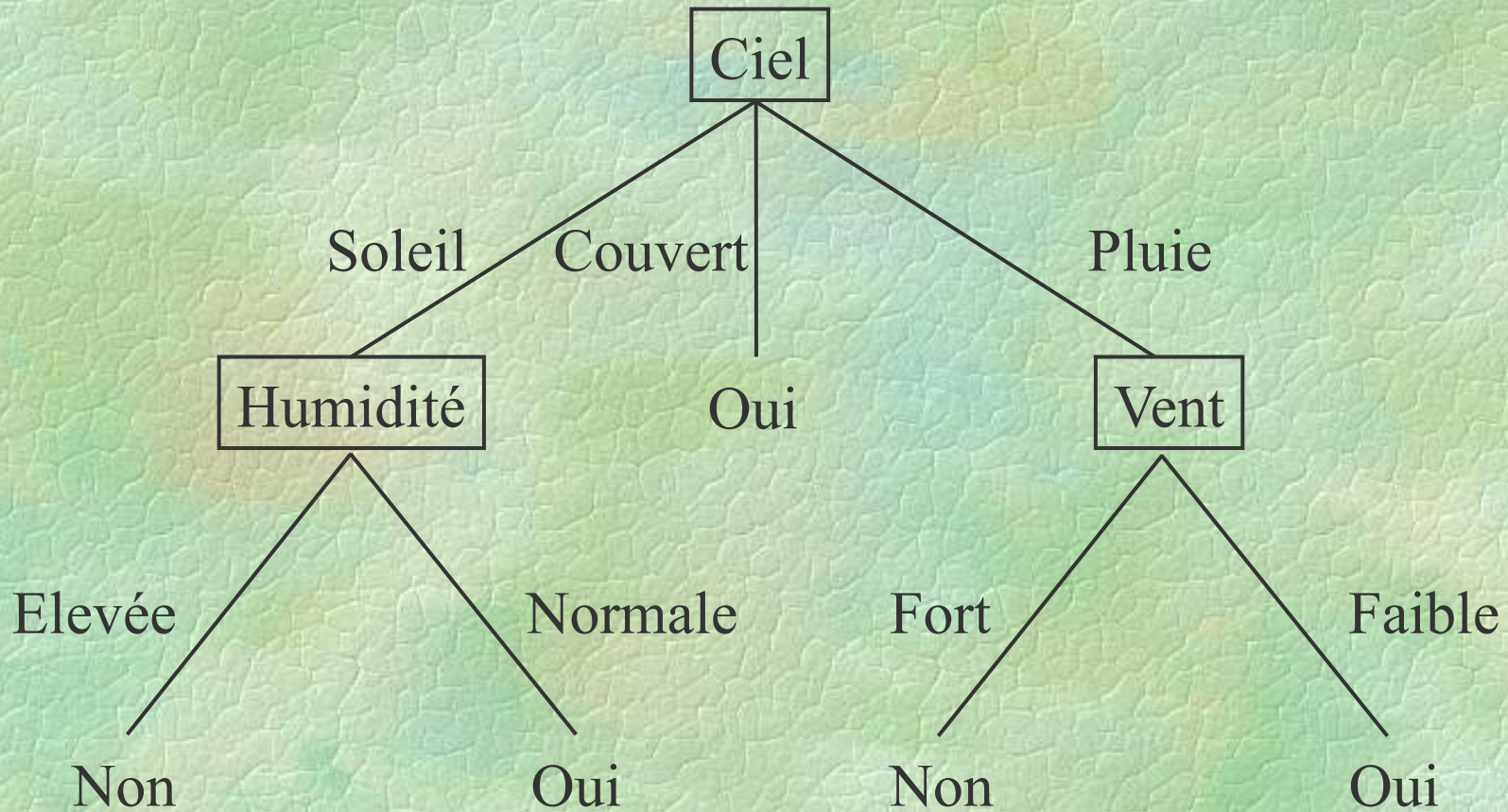


Classification d'une instance

<Ciel = Soleil, Température = Chaud, Humidité = Elevée, Vent = Fort>



Exemple d'arbre de décision



Expression sous forme logique

(Ciel = Soleil et Humidité = Normale)

ou (Ciel = Couvert)

ou (Ciel = Pluie et Vent = Faible)

Domaines d'application

- ☛ Instances représentées par des couples attribut-valeur
- ☛ Fonction cible à valeurs discrètes
- ☛ Expression disjonctive vraisemblable
- ☛ Erreurs possibles dans les exemples
- ☛ Valeurs manquantes

Médical, financier,...

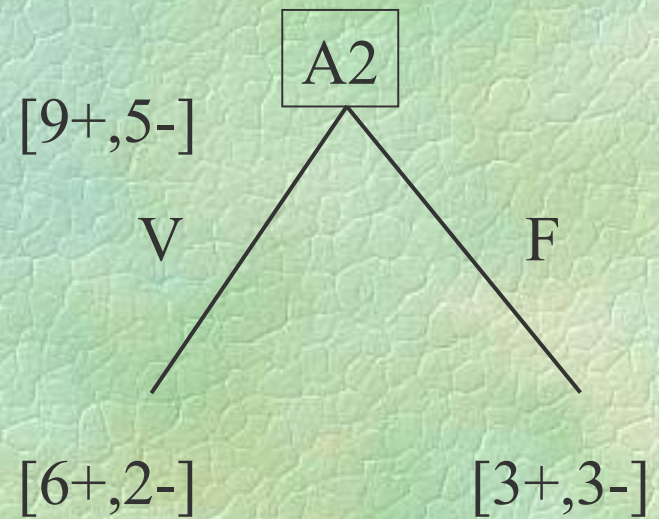
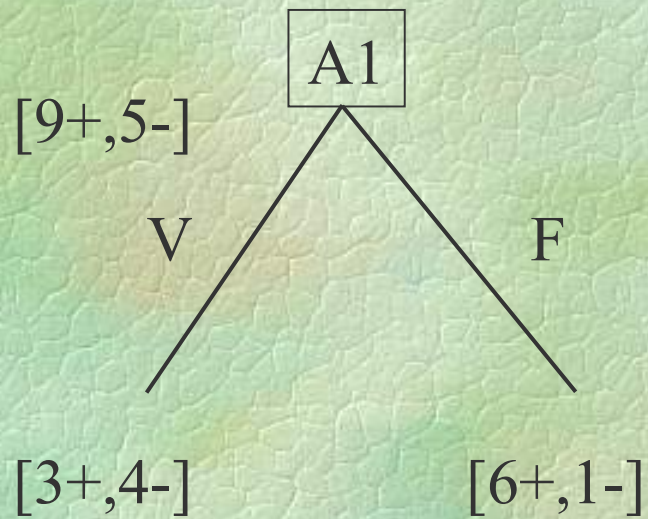
Algorithme de base

- ☛ A = MeilleurAttribut(Exemples)
- ☛ Affecter A à la racine
- ☛ Pour chaque valeur de A, créer un nouveau nœud fils de la racine
- ☛ Classer les exemples dans les nœuds fils
- ☛ Si tous les exemples d'un nœud fils sont homogènes, affecter leur classe au nœud, sinon recommencer à partir de ce nœud

Ensemble d'apprentissage

Jour	Ciel	Température	Humidité	Vent	Jouer
J1	Soleil	Chaud	Elevée	Faible	Non
J2	Soleil	Chaud	Elevée	Fort	Non
J3	Couvert	Chaud	Elevée	Faible	Oui
J4	Pluie	Doux	Elevée	Faible	Oui
J5	Pluie	Froid	Normale	Faible	Oui
J6	Pluie	Froid	Normale	Fort	Non
J7	Couvert	Froid	Normale	Fort	Oui
J8	Soleil	Doux	Elevée	Faible	Non
J9	Soleil	Froid	Normale	Faible	Oui
J10	Pluie	Doux	Normale	Faible	Oui
J11	Soleil	Doux	Normale	Fort	Oui
J12	Couvert	Doux	Elevée	Fort	Oui
J13	Couvert	Chaud	Normale	Faible	Oui
J14	Pluie	Doux	Elevée	Fort	Non

Choix de l'attribut



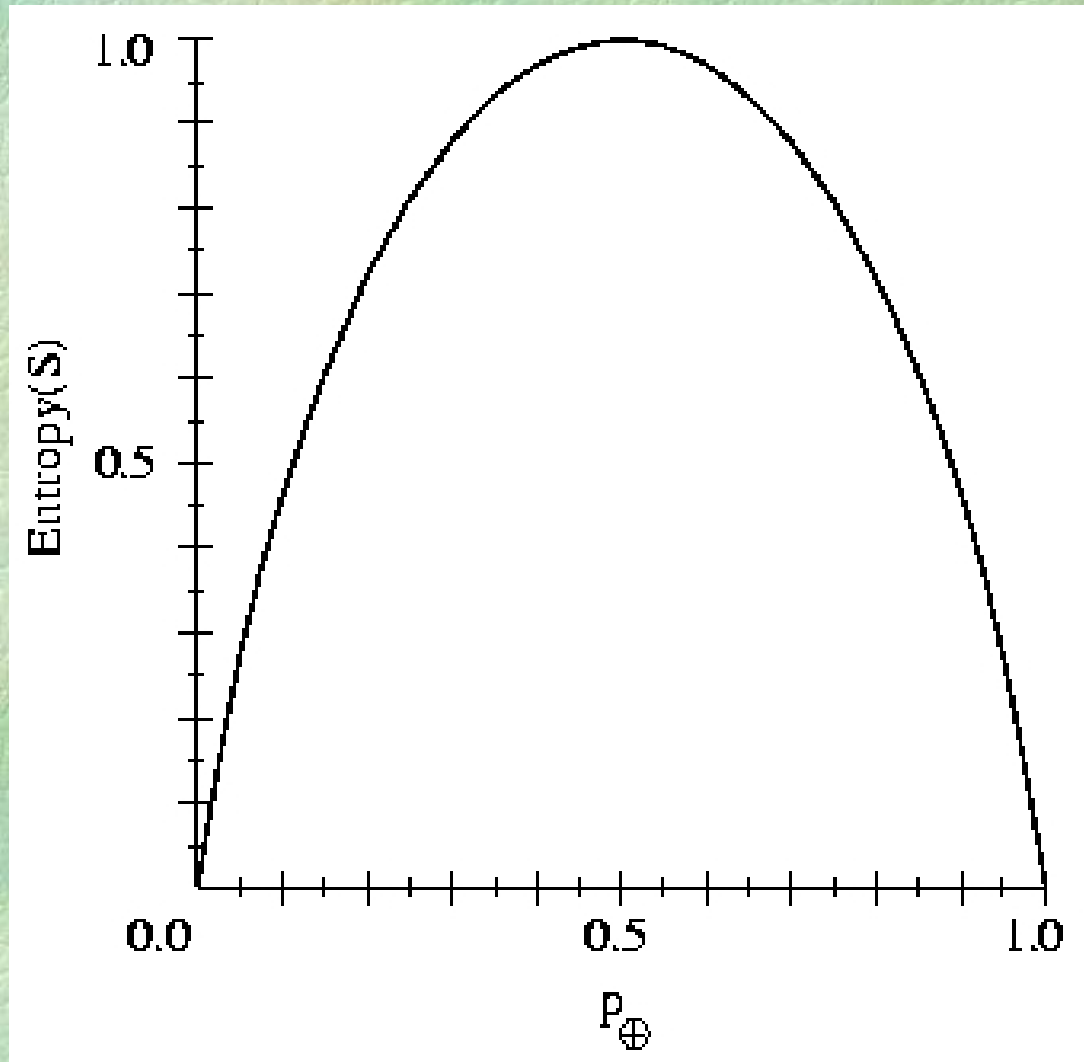
Entropie

$$\text{Entropie}(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

- ☛ S est un ensemble d'exemples
- ☛ p_+ est la proportion d'exemples positifs
- ☛ p_- est la proportion d'exemples négatifs
- ☛ Mesure l'homogénéité des exemples

$$\text{Entropie}([9+, 5-]) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940$$

Entropie



Interprétation de l'entropie

- ☛ Nombre minimum de bits nécessaires pour coder la classe d'un élément quelconque de S
- ☛ Théorie de l'information : un code de longueur optimale utilise $-\log_2 p$ bits à un message de probabilité p .

$$\textit{Entropie}(S) \equiv p_+(-\log_2 p_+) + p_-(-\log_2 p_-)$$

Gain d'information

☛ Gain(S,A)=Réduction d'entropie due à un tri suivant les valeurs de A

$$Gain(S,A) \equiv Entropie(S) - \sum_{v \in Valeurs(A)} \frac{|S_v|}{|S|} Entropie(S_v)$$

Choix de l'attribut

[9+,5-]

E=0,940

Humidité

Elevée

Normale

[3+,4-]

E=0,985

[6+,1-]

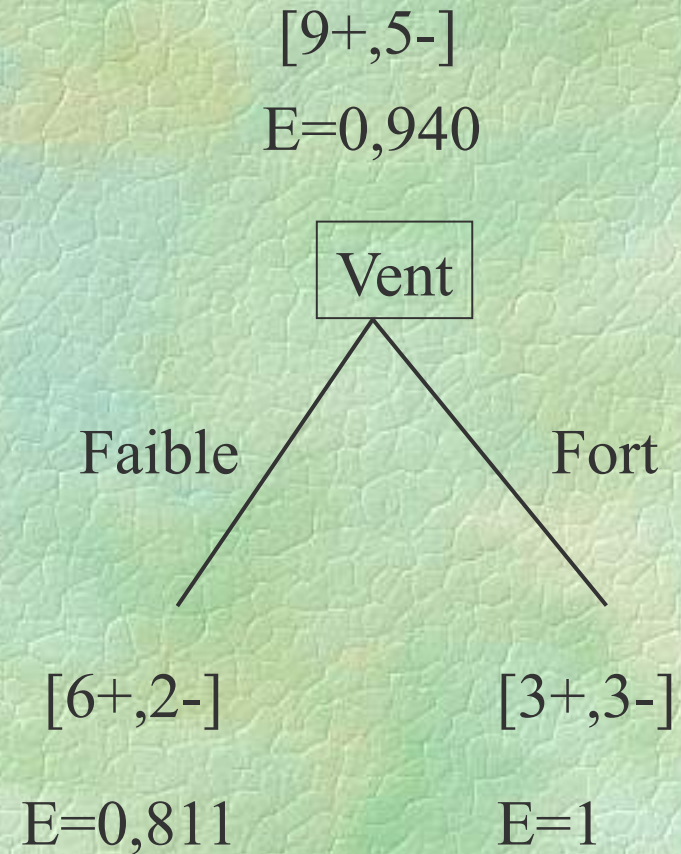
E=0,592

Gain(S, Humidité)

$$= 0,940 - (7/14)0,985 - (7/14)0,592$$
$$= 0,151$$

Choix de l'attribut

$$\begin{aligned} \text{Gain}(S, \text{Vent}) &= 0,940 - (8/14)0,811 - (6/14) 1 \\ &= 0,048 \end{aligned}$$



Choix de l'attribut

[9+,5-]

E=0,940

Ciel

Soleil

Couvert

Pluie

[2+,3-]

[4+,0-]

[3+,2-]

E=0,971

E=0

E=0,971

Gain(S, Ciel)

= 0,940 - (5/14)0,971 - (5/14)0,971 - 0

= 0,246

Choix de l'attribut

[9+,5-]

E=0,940

Température

Chaud

Doux

Froid

[2+,2-]

[4+,2-]

[3+,1-]

E=1

E=0,918

E=0,811

$$\begin{aligned} \text{Gain}(S, \text{Température}) \\ &= 0,940 - (4/14)1 - (6/14)0,918 - (4/14)0,811 \\ &= 0,029 \end{aligned}$$

Choix de l'attribut

Humidité

Elevée

Normale

$$\text{Gain}(S, \text{Humidité}) = 0,151$$

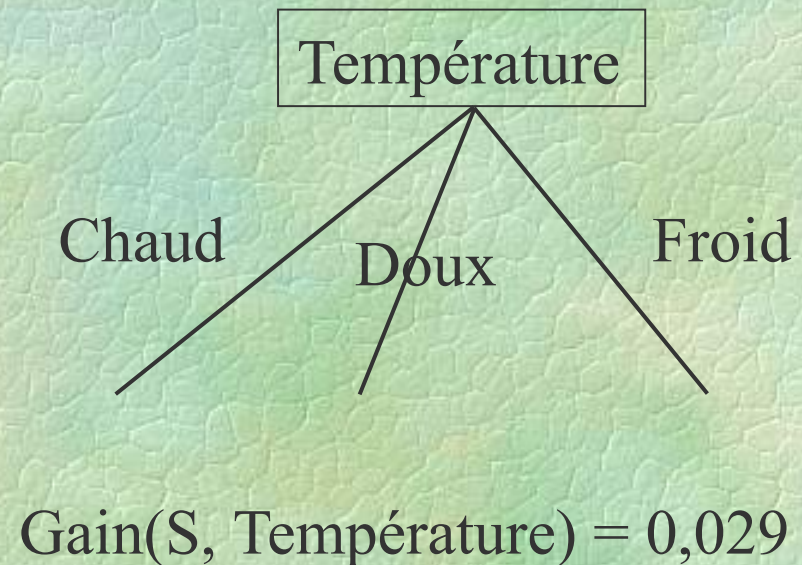
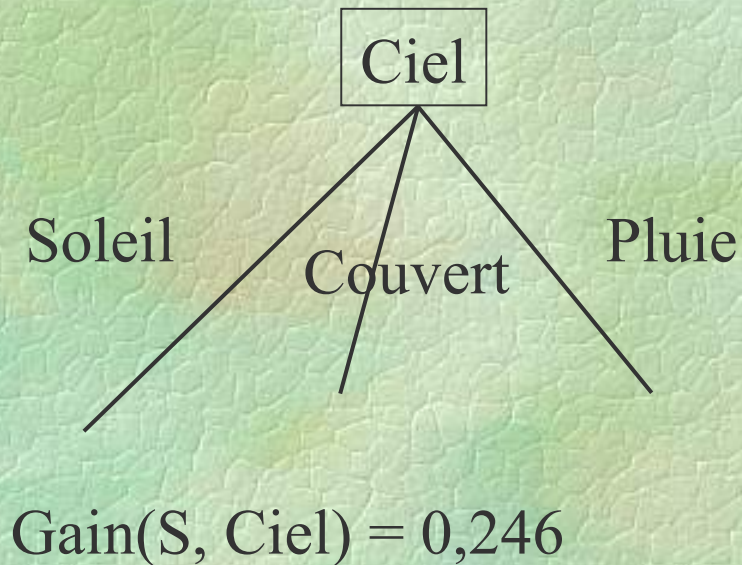
Vent

Faible

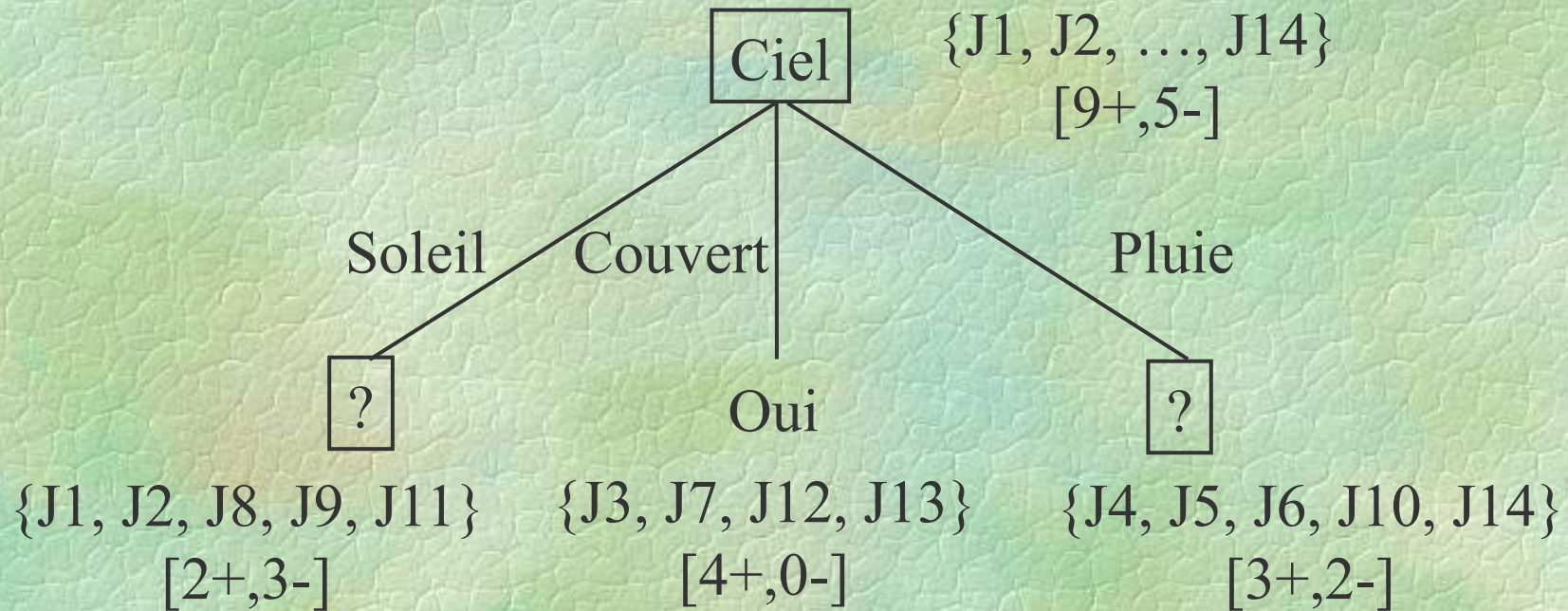
Fort

$$\text{Gain}(S, \text{Vent}) = 0,048$$

Choix de l'attribut



Choix du prochain attribut

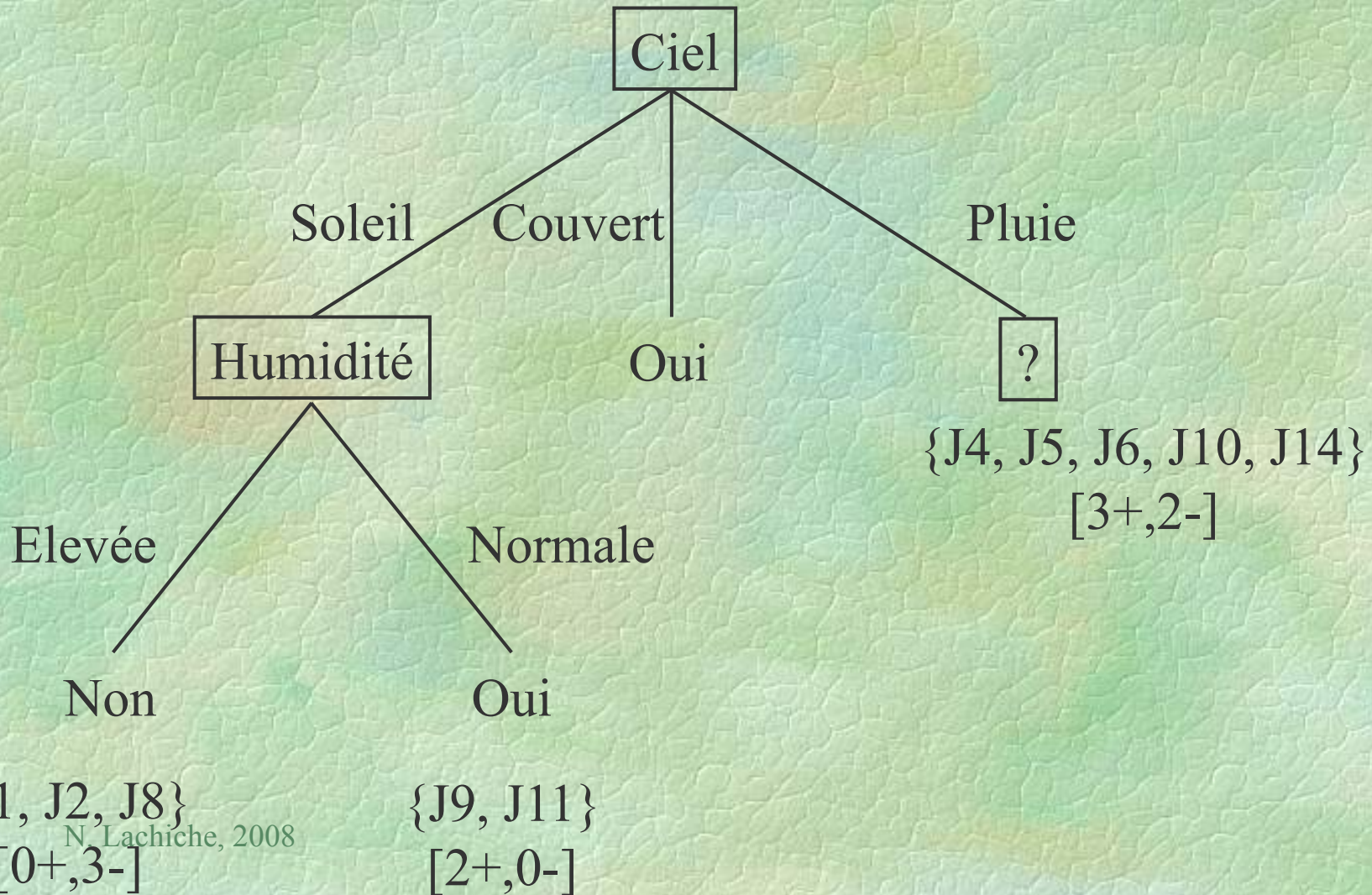


$$\text{Gain}(S_{\text{Soleil}}, \text{Humidité}) = 0,970 - (3/5) 0 - (2/5) 0 = 0,970$$

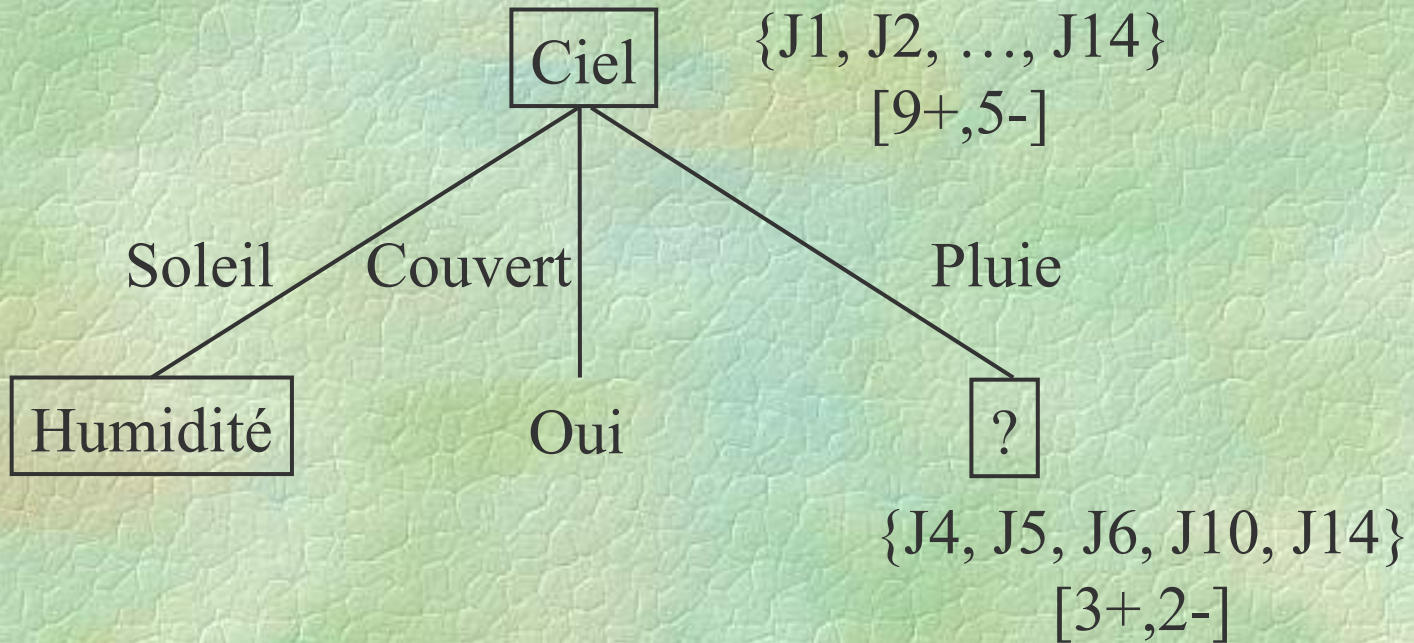
$$\text{Gain}(S_{\text{Soleil}}, \text{Température}) = 0,970 - (2/5) 0 - (2/5) 1 - (1/5) 0 = 0,570$$

$$\text{Gain}(S_{\text{Soleil}}, \text{Vent}) = 0,970 - (2/5) 1 - (3/5) 0,918 = 0,019$$

Exemple d'arbre de décision



Choix du prochain attribut

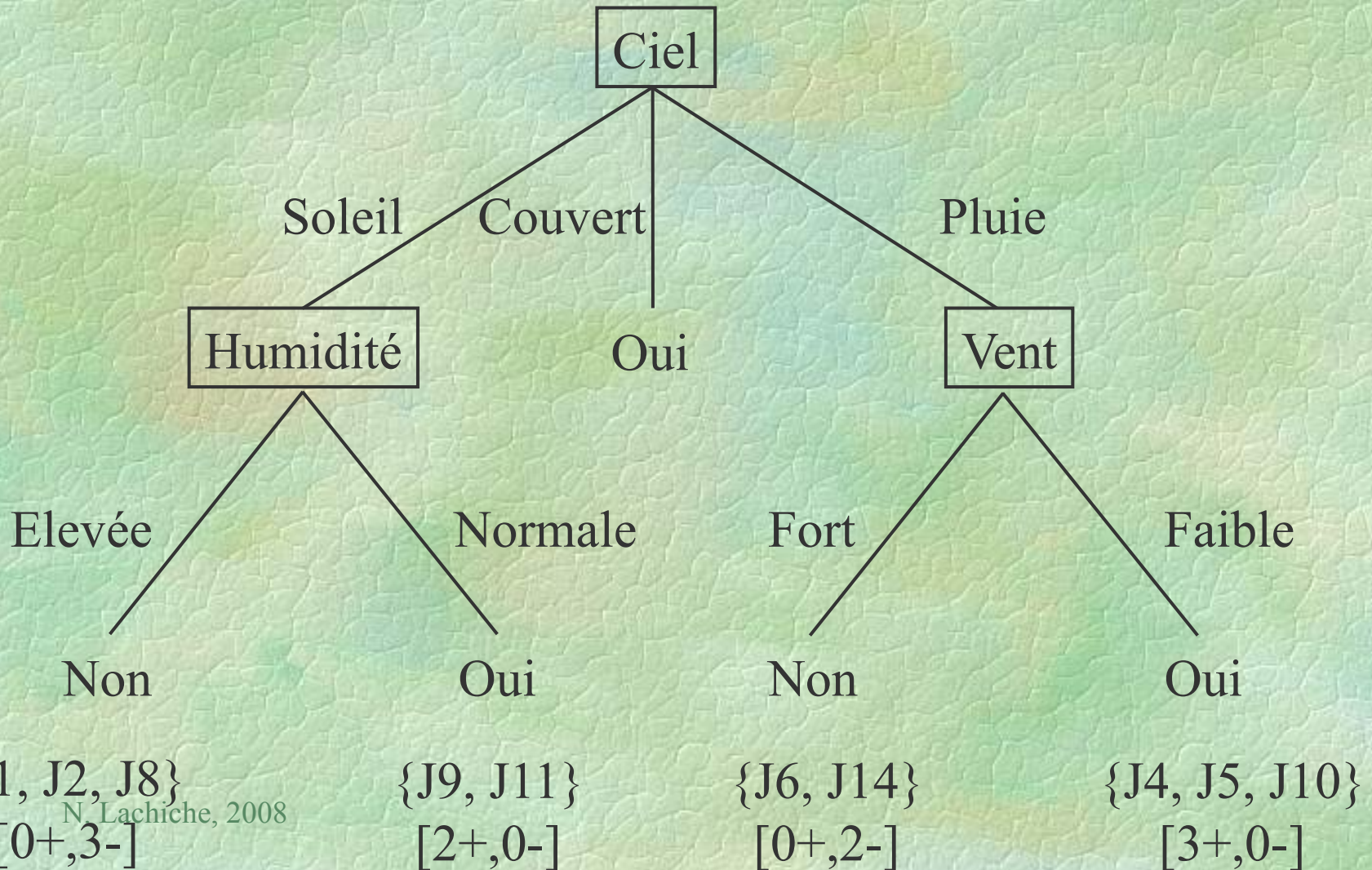


$$\text{Gain}(S_{\text{Pluie}}, \text{Humidité}) = 0,970 - (2/5) 1 - (3/5) 0,918 = 0,019$$

$$\text{Gain}(S_{\text{Pluie}}, \text{Température}) = 0,970 - (0/5) - (3/5) 0,918 - (2/5) 1 = 0,019$$

$$\text{Gain}(S_{\text{Pluie}}, \text{Vent}) = 0,970 - (2/5) 0 - (3/5) 0 = 0,970$$

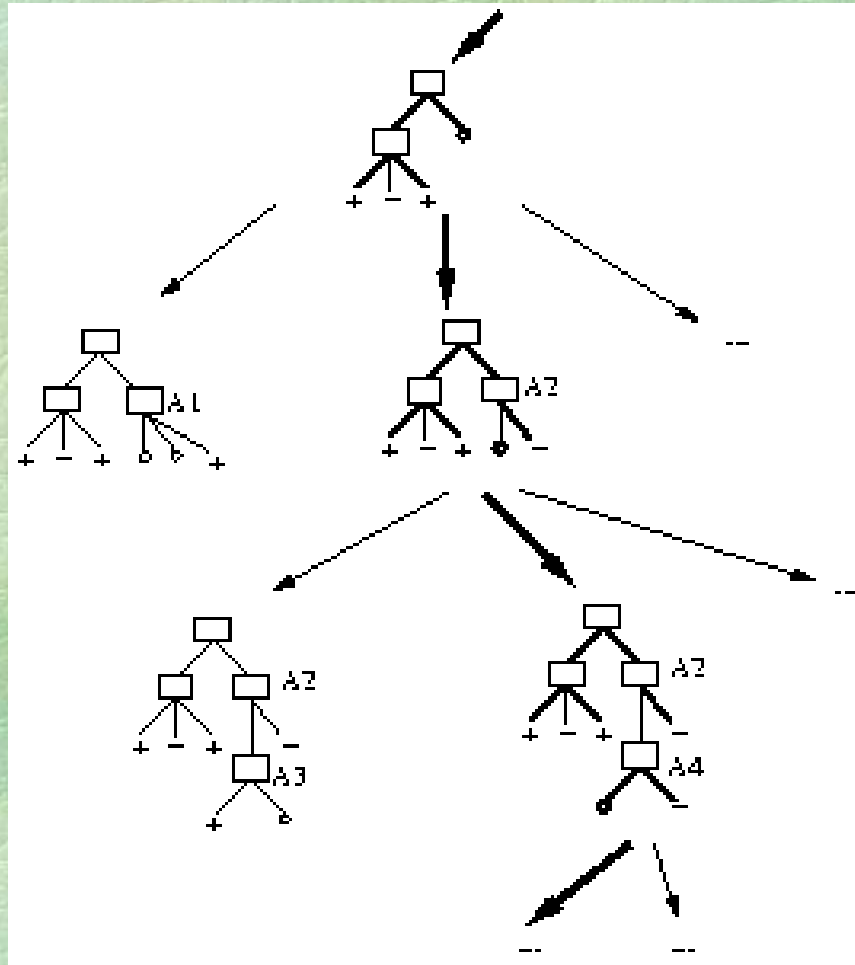
Exemple d'arbre de décision



Espace des hypothèses d'ID3

- ☛ Apprentissage vu comme une recherche dans un espace d'hypothèses
- ☛ « Hill-climbing » à partir de l'arbre vide, guidé par le gain d'information

Espace des hypothèses d'ID3



Espace des hypothèses d'ID3

- ☛ Espace des hypothèses est complet
- ☛ Rend une seule solution, pas toutes...
- ☛ Pas de retour en arrière
- ☛ Choix faits sur des critères statistiques

Biais inductif d'ID3

- ☛ « préfère les arbres les plus courts »
- ☛ ceux qui placent les attributs de meilleurs gains d'information près de la racine
- ☛ Approche heuristique d'une recherche en largeur d'abord

Types de biais

☞ Biais de restriction

- biais de langage

☞ Biais de préférence

- biais de recherche

Pourquoi préférer les hypothèses les plus courtes ?

☞ Rasoir d'Occam

- Préférer les hypothèses les plus simples qui expliquent les données

☞ Plus générale, plus de chances d'être réfutée

☞ Taille de l'hypothèse dépend du langage

Extensions

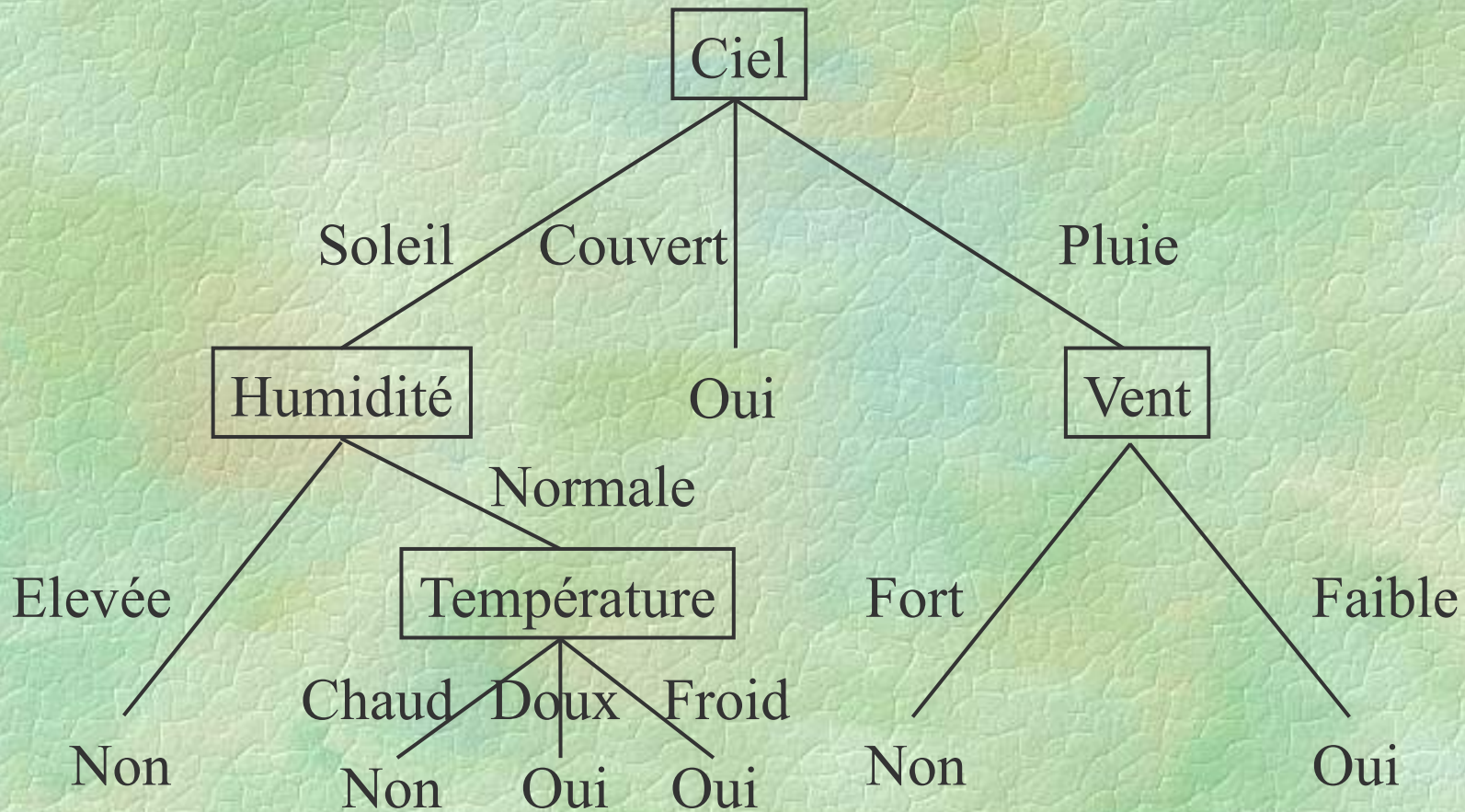
- ☛ Sur-apprentissage
- ☛ Attributs continus
- ☛ Ratio du gain d'information
- ☛ Valeurs manquantes
- ☛ Coût des attributs

Sur-apprentissage

☞ Effet des exemples bruités :

- J15 <Soleil, Chaud, Normale, Fort, Non>

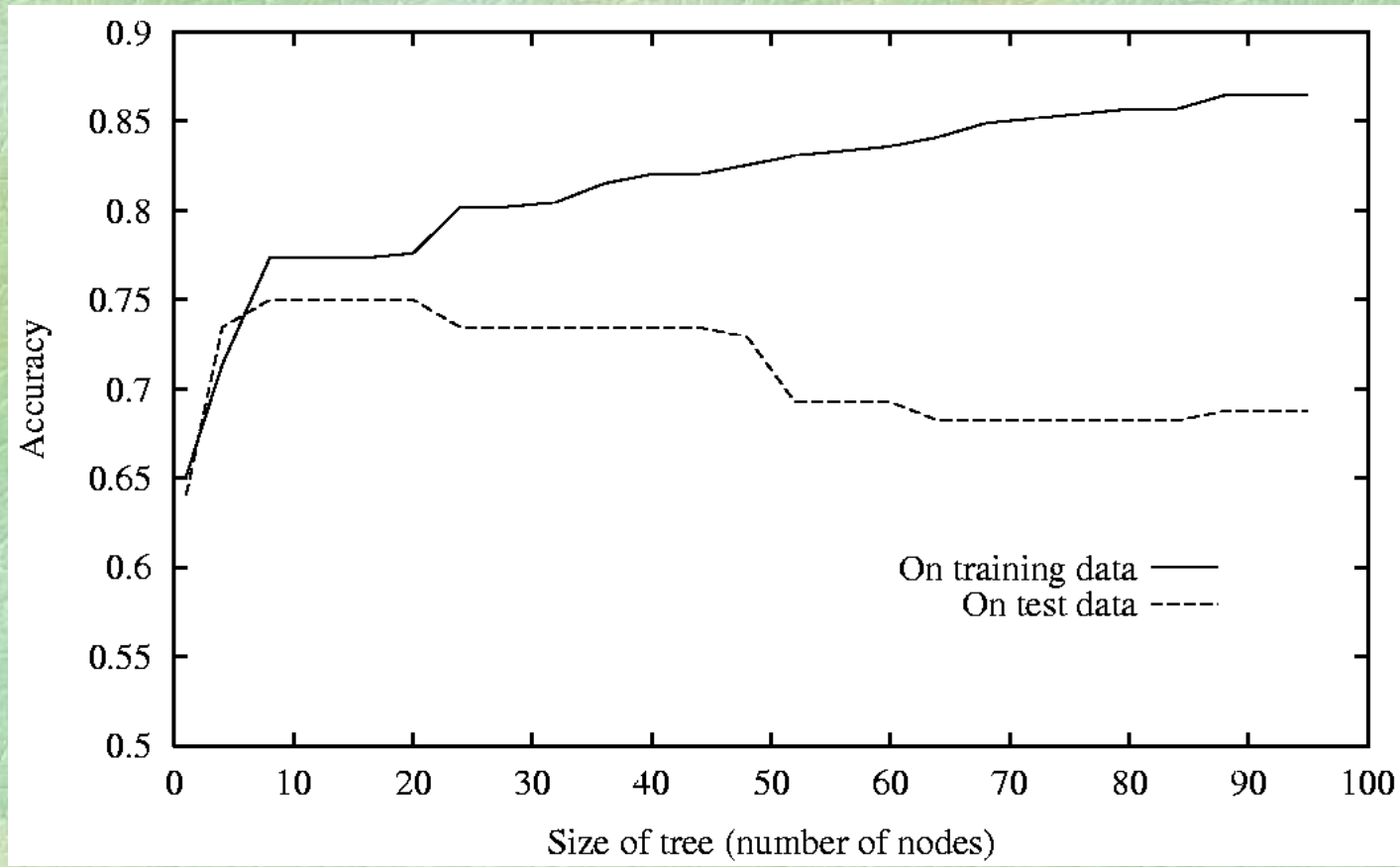
Sur-apprentissage



Sur-apprentissage

- ☛ Erreur expérimentale : $\text{erreur}_{\text{exp}}(h)$
- ☛ Erreur réelle sur la distribution D des instances : $\text{erreur}_D(h)$
- ☛ Sur-apprentissage :
 - $\text{erreur}_{\text{exp}}(h) < \text{erreur}_{\text{exp}}(h')$
 - $\text{erreur}_D(h) > \text{erreur}_D(h')$

Exemple de sur-apprentissage



Eviter le sur-apprentissage

- ☞ Arrêter la croissance de l'arbre quand la division des données n'est plus statistiquement significative
- ☞ Générer l'arbre entier, puis élaguer

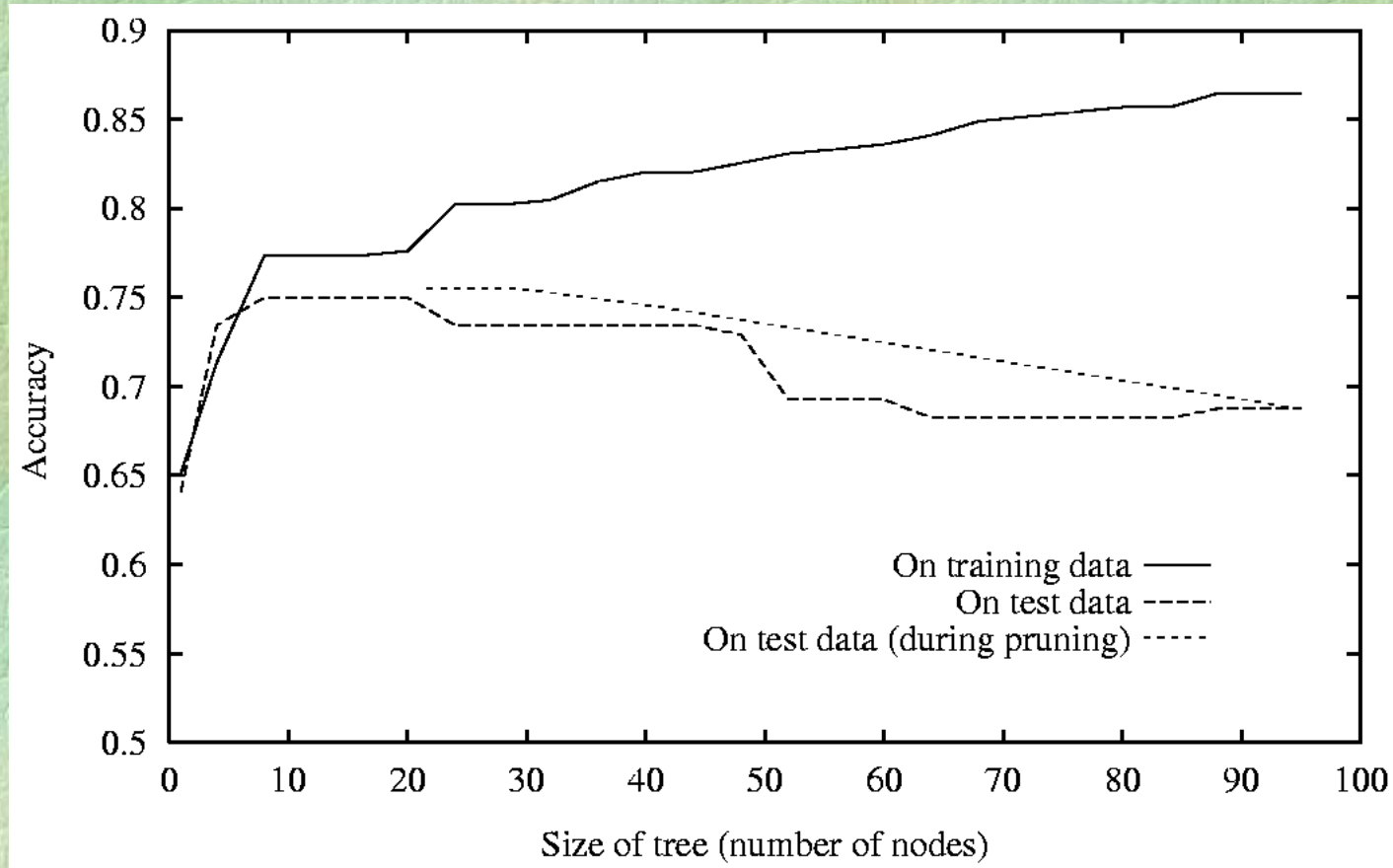
Sélection du « meilleur » arbre

- ☛ Mesurer les performances sur un ensemble distinct de données de validation
- ☛ Mesurer les performances sur l'ensemble d'apprentissage et effectuer test statistique
- ☛ MDL : minimiser $\text{taille}(\text{arbre}) + \text{taille}(\text{erreurs de classification}(\text{arbre}))$

Élagage basé sur l'erreur

- ☛ Diviser les données en ensembles d'apprentissage et de validation
- ☛ Tant que l'élagage réduit l'erreur
 - Evaluer sur l'ensemble de validation l'influence d'un élagage à partir de chaque nœud
 - Effectuer le meilleur élagage

Effet de l'élagage basé sur l'erreur



Post-élagage des règles

- ☛ Convertir l'arbre en un ensemble de règles
- ☛ Elaguer chaque règle indépendamment
- ☛ Ordonner les règles obtenues en fonction de leur précision

- ☛ Estimation de la précision d'une règle
 - ensemble de validation
 - estimation pessimiste (C4.5)

Intérêts de la conversion en règles

- ☛ Supprime la distinction entre les nœuds
- ☛ Plus flexible que l'élagage de l'arbre
- ☛ Améliore la lisibilité

Estimation pessimiste C4.5

☛ Ciel=couvert^Humidité=normale→Oui

- $n=40, r=12, \text{erreur}_S=12/40=0,3$
- $\sigma=[r/n(1-r/n)/n]^{1/2}=[0,3 \times 0,7/40]^{1/2}=0,07$
- $\text{erreur}_S+z_N \sigma = 0,3+1,96 \times 0,07=0,437$

☛ Humidité=normale→Oui

- $n=160, r=56, \text{erreur}_S=56/160=0,35$
- $\sigma=[r/n(1-r/n)/n]^{1/2}=[0,35 \times 0,65/160]^{1/2}=0,04$
- $\text{erreur}_S+z_N \sigma = 0,35+1,96 \times 0,04=0,424$

Attributs continus

☞ Créer un attribut discret

- $(\text{Température} > 25) = V, F$

Température	5	9	15	20	30	35
Jouer	Non	Non	Oui	Oui	Non	Non

Choix du seuil

☛ Entropie [2+,4-] = 0,918

☛ $\text{Gain}_1 = 0,918 - (1/6)0 - (5/6)0,970 = 0,109$

☛ $\text{Gain}_2 = 0,918 - (2/6)0 - (4/6)1 = 0,251$

☛ $\text{Gain}_3 = 0,918 - (3/6)0,918 * 2 = 0 !$

☛ $\text{Gain}_4 = 0,918 - (4/6)1 - (2/6)0 = 0,251$

☛ $\text{Gain}_5 = 0,918 - (5/6)0,970 - (1/6)0 = 0,109$

Ratio du gain d'information

- ☛ Si un attribut a beaucoup de valeurs, le gain d'information le sélectionnera
- Plus les ensembles sont petits, plus ils sont purs

$$\textit{GainRatio}(G,A) \equiv \frac{\textit{Gain}(S,A)}{\textit{SplitInformation}(S,A)}$$

$$\textit{SplitInformation}(S,A) \equiv - \sum_{i=1}^c \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Valeurs manquantes

- ☛ Si un exemple n'a pas de valeur pour A
 - si un nœud teste A, utiliser la valeur la plus commune parmi les exemples de ce nœud
 - utiliser la valeur la plus fréquente parmi les exemples de la même classe
 - affecter une probabilité à chaque valeur de A

Coût des attributs

☞ Exemples

- diagnostic médical
- robotique

☞ Heuristiques

$$\frac{Gain^2(S,A)}{Coût(A)}$$

$$\frac{2^{Gain(S,A)} - 1}{(Coût(A) + 1)^w}, w \in [0; 1]$$